ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2022 December: 28(4): 347~364 http://dx.doi.org/10.13088/jiis.2022.28.4.347

The Detection of Online Manipulated Reviews Using Machine Learning and GPT-3*

Olga Chernyaeva
College of Business Administration,
Pusan National University
(misslelka@pusan.ac.kr)

Taeho Hong
College of Business Administration,
Pusan National University
(hongth@pusan.ac.kr)

Fraudulent companies or sellers strategically manipulate reviews to influence customers' purchase decisions; therefore, the reliability of reviews has become crucial for customer decision-making. Since customers increasingly rely on online reviews to search for more detailed information about products or services before purchasing, many researchers focus on detecting manipulated reviews. However, the main problem in detecting manipulated reviews is the difficulties with obtaining data with manipulated reviews to utilize machine learning techniques with sufficient data. Also, the number of manipulated reviews is insufficient compared with the number of non-manipulated reviews, so the class imbalance problem occurs. The class with fewer examples is under-represented and can hamper a model's accuracy, so machine learning methods suffer from the class imbalance problem and solving the class imbalance problem is important to build an accurate model for detecting manipulated reviews. Thus, we propose an OpenAI-based reviews generation model to solve the manipulated reviews imbalance problem, thereby enhancing the accuracy of manipulated reviews detection. In this research, we applied the novel autoregressive language model - GPT-3 to generate reviews based on manipulated reviews. Moreover, we found that applying GPT-3 model for oversampling manipulated reviews can recover a satisfactory portion of performance losses and shows better performance in classification (logit, decision tree, neural networks) than traditional oversampling models such as random oversampling and SMOTE.

Keywords : Text Mining, Online Reviews, Manipulated Reviews Detection, Text Generation, Class Imbalance Problem.

Received : November 20, 2022 Revised : December 14, 2022 Accepted : December 15, 2022
Publication Type : Regular Track Corresponding Author : Taeho Hong

* This work was supported by a 2-Year Research Grant of Pusan National University.

1. Introduction

Online customer reviews are among the most influential resources for searching for information before purchasing a product or service and for sharing the post-purchase experience. Therefore, the amount of available online reviews grows at an exponential rate. Since online platforms customers can in real-time access and share opinions about products or services (Li et al., 2022; Dwivedi et al., 2020), utilizing online reviews written in e-commerce helps customers make better decisions such as search and purchase processes (Park et al., 2017; Cheng & Ho, 2015). According to Eslami et al. (2018), around 85% of customers read online reviews written by previous customers and, based on reviews, make a purchase decision. Also, for e-commerce, online reviews are one of the most trusted sources of information for sharing personal past purchase experiences (Salminen et al., 2022; Ott et al., 2011). Online platforms with millions of reviews, such as Yelp.com or Amazon.com, have become an important source for connecting with customers, receiving real-time feedback, and promoting new products or offers.

The study conducted by Mayzlin et al. (2014) have investigated how marketers can strategically manipulate customers' perceptions and opinions of products or services through online channels such as online reviews of retail sites. They discovered that fraudulent companies or sellers strategically manipulate reviews to influence customers' purchase decisions. Moreover, Hu et al. (2011) state that "unethical users manipulate online reviews; they can either post reviews with a high numeric rating or manipulate the textual statements posted in the review." Posting an untruthful review or a review without accounting for a real customer's experience can be considered manipulation (Tian et al., 2020). Therefore, manipulation of online reviews can occur when online vendors or agencies hired by them produce customers' reviews by posing as real customers.

To detect manipulated reviews, usually, researchers face the problem of a need for more data, such as detailed information about the reviewer, posting frequency, geolocation, and information about the seller. Only the owners of the online platforms themselves have the opportunity to receive information and create a manipulated review filtering algorithm (Crawford et al., 2015). For example, Yelp.com has an automated filtering model based on AI. In 2021, Yelp.com announced that about 22% of 19.6 million reviews were not recommended by automated recommendation software. Automated recommendation software was designed to detect conflicts of interest and manipulated low-quality or less reliable reviews based on detailed information about the reviewer that only platform owners can access, such as IP, posting frequency, geolocation, and other information about the seller. In other words, we can call not recommended reviews - manipulated reviews. However, even with a dataset with non-recommended reviews, their number is insufficient to build an accurate model for detecting manipulated reviews based on publicly available textual information. Therefore, we are facing a class imbalance problem. The class imbalance problem is common in text mining, especially classification problems. Since classes with fewer examples are under-represented and can

hamper a model's accuracy, the imbalanced problem has received attention from the academic community. However, the manipulated reviews detection field needs to gain insights into the relationships among the degree of imbalance, loss of performance, and the recovery capacity of treatment methods. Therefore, this research aims to generate reviews based on a non-recommended reviews dataset, thereby solving the class imbalance problem.

First, in this research, we explored the causal link between the data imbalance problem and manipulated reviews detection models' performance loss. We compare the results of manipulated reviews detection using different degrees of imbalanced distributions. Second, we tested the capacity of oversampling methods to recover from the loss of performance caused by imbalanced review datasets. We applied two widely used oversampling techniques: random oversampling and the synthetic minority oversampling technique (SMOTE). We compared them with generating manipulated reviews by the novel autoregressive language model - GPT-3 to generate reviews based on manipulated reviews. Third, we explore the levels of sensitivity and specificity of different detection methods (logit, decision tree, neural networks) to imbalanced datasets and the levels of sensitivity and specificity of sampling techniques to various training datasets and the test set.

2. Literature Review

2.1. Manipulated Reviews

In traditional business, the manipulation of news or tweets is not a new area of research (Majumdar et al., 2007). However, review manipulation is one of the latest and crucial issues in the e-commerce service area. For example, on online retail sites, 10.3% of existing online reviews were manipulated (Hu et al., 2012). However, until recently, even prominent online vendors like Yelp.com, Amazon.com, and TripAdvisor.com rarely discussed how online retail sites should fight online customer reviews manipulations. Before, vendors never disclosed the use of unethical users who create fraud reviews because these unethical users could take advantage of this knowledge (Chen & Lin, 2013).

Not all companies can collect the best online reviews and be rated the highest. Some vendors develop strategies to control customer opinion (Gössling et al., 2016) when they seek to expand the online customer base or when online evaluations threaten the customer base. This strategy to control customer opinion to influence a vendor's reputation and attract new customers takes various forms, including improved services and manipulation (Banerjee & Chua, 2014). According to Anderson and Simester (2014) and Filieri (2015), "fake reviews" were identified as one form of manipulation. Secret content control algorithms, which can identify and specify forms of manipulation, review, or rating manipulations, have even appeared on famous platforms such as Amazon.com and TripAdvisor.com (Weisberg et al., 2011). Since online reviews directly impact purchase decision-making, ultimately affecting the sale of products and services (Cao et al., 2011), online reviews remain essential and play a central role in purchasing decisions. Therefore, detecting

manipulation in online customer reviews is critical for e-commerce.

2.2. Manipulated Reviews Detection

E-commerce and online platforms are interested in receiving authentic feedback from customers that can help improve products' quality and satisfaction rate. However, manipulated reviews can risk the credibility and image of the platform, especially in the case of online platforms such as Yelp.com or Amazon.com (Ismagilova et al., 2020). Since online retail sites use online reviews to determine a product's ranking among other products in the same category, online customer reviews are the most powerful instrument of e-commerce (Gobi & Rathinavelu, 2019). Therefore, some unethical companies can manipulate reviews to achieve their purpose, for example, damaging the reputation of competitors. Moreover, the negative impact of manipulated reviews can lead not only to reputation harm but also lead to financial costs. For instance, by decreasing only one star, the rating on Yelp.com cost a 5-9% decrease in revenue (Luca, 2011). As a result, when the rating is inflated or deflated through reviews, the market faces unfair competition (He et al., 2022). Therefore, to protect companies from unfair competition, it is important to understand the nature of creating manipulated reviews before detecting manipulated reviews.

Manipulated reviews can be created in two main ways: human-generated and computer-generated (Salminen et al., 2022). In the case of human-generated reviews, the author never uses a product or service but creates a review for a fee. It can be scaled in a "market of fakes," where an unethical company orders to create positive or negative manipulated reviews in a given quantity (He et al., 2022). In the case of computer-generated, a company uses text-generation techniques to create manipulated reviews such as natural language processing and machine learning.

Since there are a lot of theories and methods to detect manipulation (Tsikerdekis & Zeadally, 2014), researchers are motivated to develop and study more sophisticated methods. The studies on manipulation detection techniques can be generally classified into three research techniques: Machine learning is a subset of artificial intelligence (AI). Machine learning algorithms widely use sample data (training data) based on a mathematical model to predict or classify. In non-machine learning cases, this research does not have training data to provide a statistical model.

Supervised learning-based manipulation detection techniques and unsupervised learning-based manipulation detection techniques are built based on the principles of design science, machine learning techniques, and verbal and nonverbal features to detect manipulation (Nunamaker et al., 2016). In literature about OCRs manipulation detection, supervised learning is the most common method used for OCRs manipulation detection and examines labeled data's learning. However, providing supervised learning labeled data is required to train a classifier, which presents a challenge in review manipulation detection. In other words, in the case of detecting products with manipulated online reviews, the primary condition for carrying out supervised learning methods is the availability of a dataset with fake online reviews (Lim et al., 2010).

〈Table 1〉 Previous Studies Detected Manipulated Reviews

Method	Studies
Logistic regression	Banerjee & Chua, 2014; Ho et al., 2016; Khurshid et al., 2019; Liu et al., 2019
K-nearest neighbors	Rajamohana & Umamaheswari, 2018; Rajamohana et al., 2017
Random forest	Jalther and Priya, 2019; Zhang et al., 2016
Decision tree	Ball and Elworthy, 2014, Khurshid et al., 2019
Neural networks	Li et al., 2017; Ren and Ji, 2017

〈Table 2〉 Previous Studies Applied Oversampling Methods

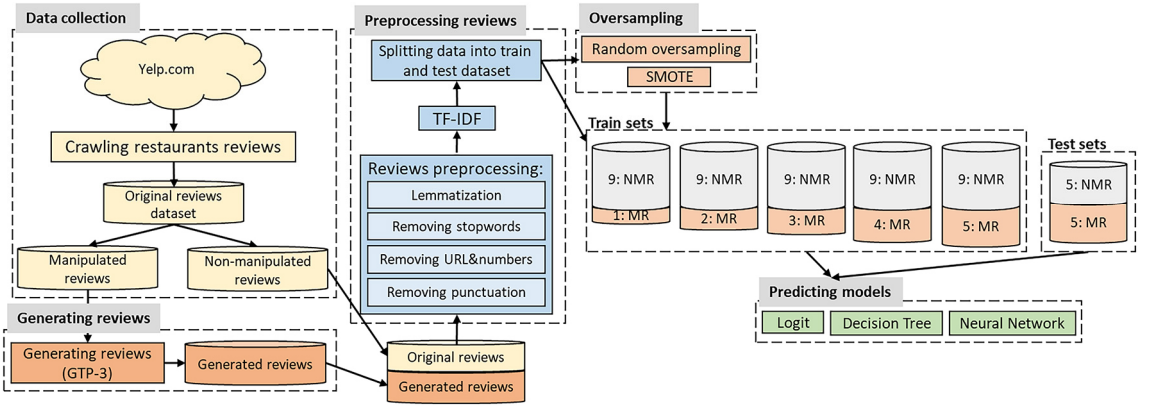
Study	Oversampling Methods	Description
Li et al., 2013	Random Oversampling	Prediction of default risk
Veganzones & Séverin, 2018	Random Oversampling, Synthetic minority Oversampling, SMOTE	Bankruptcy prediction
Douzas et al., 2018	SMOTE, k-mean SMOTE	Train classification algorithms
Mouratidis et al., 2021	SMOTE	Fake news detection
Salminen et al., 2022	GPT-2, ULMFiT	Fake reviews detection

A list of studies that examined supervised learning techniques to detect manipulated reviews is shown in Table 1. However, these studies did not solve the class imbalance problem of the review dataset. Since the number of manipulated is lower than non-manipulated reviews, primarily solving the class imbalanced problem is critical to building an accurate model for detecting manipulated reviews.

2.3. The Class Imbalanced Problem in Manipulated Reviews Detection

The class imbalance problem occurs when one class is under-represented, which can hamper a model's accuracy (Kotsiantis et al., 2006). The class imbalance problem is common in the classification field because the classification models aim to optimize the overall accuracy of class prediction (López et al., 2013). However, these models commonly do not consider the unequal distributions between classes that lead to the underrepresentation of the data characteristics of an imbalanced class (Kim et al., 2015). As a result, because of biasing toward the majority class, the classification model with an imbalanced class problem accurately classifies the majority class and misclassifies the minority class (Kim & Kwahk, 2022; Fernández et al., 2010). Therefore, the imbalanced problem has received attention from the academic community.

A list of previous existing studies that had oversampled minority classes to solve the imbalanced class problem is shown in Table 2. Oversampling methods such as random oversampling, synthetic minority oversampling, and SMOTE are common methods to oversample numeric data for predicting



〈Figure 1〉 Research Framework for Detecting Manipulated Online Reviews.

bankruptcy and default risk and for training classification models.

In the case of manipulated review detection, the classes of manipulated and non-manipulated reviews usually are highly imbalanced. Therefore, to create an efficient manipulated reviews detection model solving the imbalanced class problem is necessary. In previous studies, researchers applied two methods to oversample reviews. First, to oversample textual data, text data were transformed into a numerical format using natural language processing (NLP) techniques such as TF-IDF, word embedding, etc. (Suh et al., 2017). Second, to solve the class imbalance problem between fake and real reviews, previous researchers applied GPT-2 and ULMFiT. According to Salminen et al. (2022), GPT-2 showed better performance than ULMFiT in generating reviews. However, previous studies did not compare these two methods, and therefore in our study, by comparing the results of the two methods above, we want to determine which method is effective in detecting manipulated reviews. Moreover, in this study, the latest version of GPT-3, upgraded from GPT-2, will be used.

3. Research Framework and Experiment

As shown in Figure 1, our research framework consists of five parts: data collection, generating reviews, preprocessing reviews, oversampling, and predicting models. Firstly, we collected review data from Yelp.com in the data collection step. Yelp.com publish user-created reviews about local businesses. Yelp.com introduced automated recommendation software that filters reviews into recommended and non-recommended reviews based on detailed information about the reviewer, such as IP, posting frequency, geolocation, and other information about the seller. Therefore Yelp.com is a common source of manipulated reviews in manipulated reviews detection (Kumar et al., 2022).

〈Table 3〉 Examples of manipulated and non-manipulated reviews

Review	Label
Was not happy with the lobster it was to dry. I got it for my mom cause she was caving for it. And when I asks her ther other day she told me she throw the other half away. It hurt me that she was upset.	Non-manipulated
Having worked in hospitality I usually would never leave a nasty review and I have a lot of patience for hospitality as I worked in the industry, tonifht was possibly the worst service / food I have ever received, the food was terrible and made me feel ill after it, the service was slow and lethargic.	Manipulated

In our research, we call recommended reviews as non-manipulated reviews and non-recommended reviews as manipulated reviews. An example of a non-manipulated and manipulated review is shown in Table 3. As you can see from Table 3, for humans, based only on the context of the reviews is hard to classify which review is manipulated or not.

For collecting target data, we have chosen the top restaurants in New York, the United States. The total number of reviews is 10,000 reviews. After we constructed two review datasets with manipulated (1,800 reviews) and not-manipulated reviews (8,200 reviews), preprocessing procedures were applied to the review datasets. Next, to evaluate the performance of prediction methods in imbalanced datasets, we composed the training and test sets with ratios of non-manipulated and manipulated reviews as follows:

- train set: 7,200 non-manipulated reviews, 800 manipulated reviews (9:1)
- test set: 1,000 non-manipulated reviews, 1,000 manipulated reviews (5:5)

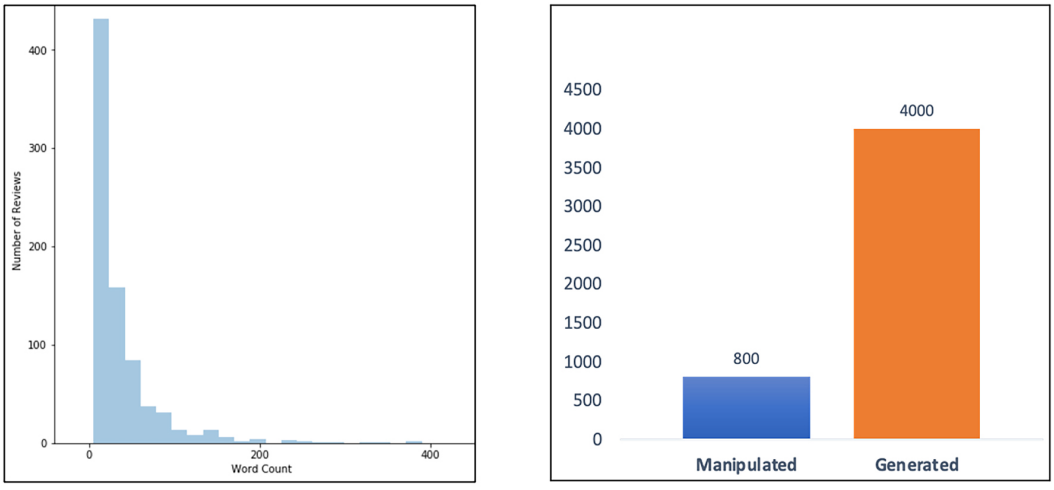
Since the data imbalance issue originates in the learning phase, we used two methods: 1) transformation of review text into a numerical format using term frequency – inverse document frequency (TF-IDF) technique and oversampling manipulated review class by applying Random Forest and SMOTE. 2) increasing manipulated reviews class examples by generating manipulated reviews using GPT-3. GPT-3 (Generative Pre-trained Transformer) is a third-generation autoregressive language model that uses deep learning created by OpenAI to generate human-like text. It uses 175 billion parameters and was trained on Azure's AI supercomputer (Scott, 2020).

According to the study by Liang & Zhu (2018), the initial number of words for effective text-generation tasks equals 5 words. Therefore, in our study, we set as input number of words equal to 5 respectively. Moreover, based on the distribution of the word count in the manipulated reviews dataset, we create discrete buckets with one-word intervals in the range of 10~200 words that were adopted as the target length for the generated reviews. An example of review-generating code in Python using GPT-3 is shown in Figure 2, where 'prompt' is input words, and 'max_length' is the length of generated review.

The word count of the reviews follows the distribution in the manipulated reviews dataset, which is shown in Figure 3 (a). As a result, we generated 4'000 reviews, distribution of original manipulated

```
In [45]: prompt = "This menu is fine"

In [46]: prompt_ids = tokenizer.encode(prompt)
         inp = tensor(prompt_ids)[None].cuda()
         inp.shape

Out[46]: torch.Size([1, 4])

In [47]: preds = learn.model.generate(inp, max_length=50, num_beams=5, temperature=0.75, early_stopping=

In [48]: tokenizer.decode(preds[0].cpu().numpy())

Out[48]: "This menu is fine, but it's not the best I've ever had.\n\nThe food is good but not as good
         as it used to be.I've been coming here since I was a kid, and I have to say that I"
```

〈Figure 2〉 A code example of review generating using GPT-3



〈Figure 3〉 (a) The word count of the reviews follows the distribution in the manipulated reviews dataset. (b) distribution of original manipulated reviews, and generated manipulated reviews in the training dataset

reviews, and generated manipulated reviews in the training dataset shown in Figure 3 (b). We increased the training dataset with non-manipulated and manipulated reviews (+ generated reviews) of 9/2, 9/3, 9/4, and 9/5, respectively.

In the last step, we detected manipulated reviews using prediction models: logit, decision tree, and neural networks and compared the results and performances of each model. Since the accuracy rate is inappropriate for imbalanced datasets because it does not account for sample distribution and focuses on predicting only the majority class, it can lead to erroneous conclusions. Therefore, to estimate and compare our models, we selected evaluation metrics that are not sensitive to

sample distribution: sensitivity, specificity, and G-mean. They are true positive (TP), true negative (TN), false negative (FN), and false positive (FP) (Shmueli et al., 2011).

$$Sensitivity = \frac{TP}{TP + FN}$$

Sensitivity is the percentage of manipulated reviews correctly detected. Sensitivity = (Number of manipulated reviews correctly detected)/(Number of all manipulated reviews).

$$Specificity = \frac{TN}{TN + FP}$$

Specificity is the percentage of non-manipulated reviews correctly detected. Specificity = (Number of non-manipulated reviews correctly detected)/ (Number of all non-manipulated reviews).

$$G - mean = \sqrt{\frac{TP}{TP + FN}} * \frac{TN}{TN + FP}$$

G-mean measures overall prediction in terms of a ratio of sensitivity and specificity.

4. Experiment Results and Analysis

To estimate model parameters, we increased the training dataset with non-manipulated and manipulated reviews (+ generated reviews) in step-by-step order 9:2, 9:3, 9:4, and 9:5. However, to estimate the model manipulated reviews detection efficiency, the number of reviews in non-manipulated and manipulated reviews datasets was equal (manipulated reviews datasets did not include generated reviews). This experimental study explores the effect of various degrees of imbalanced training sets on manipulated reviews detection. Moreover, we evaluated the performance of several oversampling techniques and compared how the manipulated reviews detection efficiency changed depending on the oversampling method. Our results demonstrate the importance of the data imbalance problem and highlight its implications for the performance of manipulated review detection. Most notably, we find that oversampling manipulated reviews can significantly improve the performance of manipulated review detection. As you can see from tables 4, 5, and 6, the model's performance has increased after applying oversampling or review generating.

Table 4 shows that the sensitivity increased after oversampling the manipulated reviews, which means that the prediction performance for manipulated reviews increased depending on the proportion of manipulated reviews in the training dataset. However, Table 5 shows that after oversampling the manipulated reviews, the specificity decreased, which means that the prediction performance for non-manipulated reviews decreased respectively. These results occur since, in the original imbalanced training set, the classification boundaries of the majority class (non-manipulated reviews) invaded the minority class (manipulated reviews), and the classification was biased toward the majority class of non-manipulated reviews. We find that sensitivity achieves a maximum rate of 45.4% and G-mean achieves a maximum

〈Table 4〉 Sensitivity rates achieved with prediction models

	Logit		Decision Tree		NN	
	Train	Test	Train	Test	Train	Test
Training set: 9:1 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Original dataset	7.6%	4.5%	22.4%	12.3%	98.6%	33.2%
Training set: 9:2 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	49.2%	19.6%	36.1%	17.8%	99.5%	38.1%
SMOTE	58.9%	24.0%	15.4%	5.1%	99.9%	39.2%
GPT-3	55.1%	12.9%	79.4%	21.7%	99.3%	37.6%
Training set: 9:3 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	91.2%	31.1%	91.2%	31.1%	99.1%	36.8%
SMOTE	75%	34.4%	97.1%	36.4%	99.8%	39.3%
GPT-3	73.6%	16.2%	99.5%	30.6%	95.7%	39.6%
Training set: 9:4 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	79.8%	38.5%	79.8%	38.5%	99.6%	36.2%
SMOTE	83.3%	40.3%	99.9%	35.8%	99.9%	42.1%
GPT-3	81.2%	17.5%	99.7%	30.7%	97.3%	41.6%
Training set: 9:5 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	85.6%	40.9%	82.4%	41.3%	99.7%	37.2%
SMOTE	87.4%	42.7%	75.9%	42.6%	99.9%	39.9%
GPT-3	85.1%	19.1%	99.7%	32.1%	98%	**45.4%**

rate of 63.4%, when the balanced proportion of the training set is 9:5, with oversampling by GPT-3 review generation and applying neural networks as a classifier.

As a result, we find that all sampling techniques improved the efficiency of manipulated reviews detection depending on the proportion of manipulated reviews in the training dataset. However, oversampling by GPT-3 review generation overperformed traditional oversampling methods such as random oversampling and SMOTE. Notable that in the case of random oversampling, sensitivity and G-mean achieved a maximum rate of 38.1% and 60.2% when the balanced proportion of the training set is 9:2 with applying neural networks as a classifier. However, in the case of SMOTE, sensitivity and G-mean achieved a maximum rate of 42.1% and 62.5% when the balanced proportion of the training set is 9:4 with applying neural networks as a classifier. In other words, our research showed that the data imbalance problem plays an important role in manipulated review detection. Also, oversampling manipulated reviews can significantly

〈Table 5〉 Specificity rates achieved with prediction models

Training set: 9:1 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
	Logit		Decision Tree		NN	
	Train	Test	Train	Test	Train	Test
Original dataset	99.9%	99.9%	99.8%	99.0%	99.9%	95.4%
Training set: 9:2 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	99.4%	99.4%	97.9%	95.8%	99.9%	95.1%
SMOTE	99.6%	99.3%	99.7%	98.1%	99.9%	94.1%
GPT-3	99.9%	99.9%	99.8%	99.0%	99.9%	95.4%
Training set: 9:3 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	98.3%	97.5%	98.3%	97.5%	99.9%	95.0%
SMOTE	98.7%	97.4%	99.5%	88.6%	99.9%	94.9%
GPT-3	99.6%	98.9%	100%	94.7%	99.9%	90.5%
Training set: 9:4 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	96.6%	95.2%	96.6%	95.2%	99.9%	94.1%
SMOTE	97.7%	95.9%	99.9%	91.0%	99.9%	92.7%
GPT-3	98.7%	97.9%	100%	89.7%	96.4%	90.1%
Training set: 9:5 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	95.1%	92.7%	84.4%	84.0%	99.9%	94.1%
SMOTE	96.6%	92.3%	84.0%	84.7%	99.9%	94.7%
GPT-3	98.5%	97.1%	100%	91.2%	94.7%	**88.5%**

improve the performance of manipulated review detection.

5. Conclusion and Further Research Direction

In this study, we have proposed an approach for manipulated reviews detection that enhanced the accuracy of manipulated reviews detection by solving the manipulated reviews imbalance problem. Furthermore, we proved that the model's performance increased after applying oversampling by review generating. In our study, we used the novel autoregressive language model - GPT-3 to generate reviews based on manipulated reviews. Moreover, we found that applying GPT-3 model for oversampling manipulated reviews can improve the sensitivity of the model to detect manipulated reviews and shows better performance in classification than traditional oversampling models such as random oversampling and SMOTE. Furthermore, the GPT-3

〈Table 6〉 G-mean values achieved with prediction models

	Logit		Decision Tree		NN	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Training set: 9:1 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Original dataset	27.6%	21.2%	47.2%	34.9%	99.3%	56.2%
Training set: 9:2 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	69.9%	44.1%	59.5%	41.3%	99.7%	60.2%
SMOTE	76.6%	48.8%	39.2%	22.4%	99.9%	60.7%
GPT-3	79.1%	35.7%	89.1%	45.3%	99.6%	58.3%
Training set: 9:3 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	83.0%	55.1%	83.0%	55.1%	99.5%	59.1%
SMOTE	86.1%	57.9%	98.3%	56.8%	99.9%	61.1%
GPT-3	85.4%	39.8%	99.8%	52.8%	96.6%	59.7%
Training set: 9:4 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	87.8%	60.5%	87.8%	60.5%	99.8%	58.4%
SMOTE	90.2%	62.2%	99.9%	57.1%	99.9%	62.5%
GPT-3	89.5%	41.3%	99.8%	52.5%	96.9%	61.2%
Training set: 9:5 non-manipulated reviews and manipulated reviews **Testing set:** 5:5 non-manipulated reviews and manipulated reviews						
Random Oversampling	90.2%	61.6%	83.3%	58.9%	99.8%	59.2%
SMOTE	91.9%	62.8%	79.8%	60.1%	99.9%	61.5%
GPT-3	91.5%	43.1%	99.8%	54.1%	96.3%	**63.4%**

based review generation model outperformed other sampling techniques on proportion level 9:5 when neural networks were applied.

During several experiments, we have: (1) explored the causal link between the data imbalance problem and manipulated reviews detection models' performance loss; (2) compared the results of manipulated reviews detection using different degrees of imbalanced distributions; (3) applied two widely used oversampling techniques: random oversampling and the synthetic minority oversampling technique (SMOTE) and compared them with generating manipulated reviews by the novel autoregressive language model - GPT-3 to generate reviews based on manipulated reviews; (4) explored the levels of sensitivity and specificity of different detection methods (logit, decision tree, neural networks) to imbalanced datasets and the levels of sensitivity and specificity of sampling techniques to various training datasets and the test set. We proved that the model's performance increased after applying

oversampling by review generating. And model's sensitivity and G-mean values increased too. Moreover, all sampling techniques achieve a recovery. However, the GPT-3 based review generation model outperformed other sampling techniques on proportion level 9:5 when neural networks were applied.

Our academic contribution consists of proposing an approach to manipulated reviews detection using machine learning techniques that improved detection accuracy and made detection more sensitive to manipulated reviews. Also, we proved that solving the imbalanced class problem of manipulated reviews dataset may significantly improve model accuracy. Moreover, we demonstrated that the OpenAI-based manipulated reviews-generating model overperformed the well-known oversampling method such as random sampling and SMOTE.

For practical contribution, our model can help general customers and retailers or service owners to distinguish manipulated reviews, thereby protecting themselves from fraudulent activities and supporting the fair competition. In addition, our model could be applied to huge datasets in which humans cannot process and identify manipulated reviews.

The limitations of our research are that we used data only from Yelp.com, the further researchers may try to analyze reviews from different platforms and not only restaurant reviews but also other categories of products or services too. Also, there are technical aspects of text generation that could benefit from future experiments, not only for fake reviews generation.

References

Anderson, E. T., & Simester, D. I. (2014). Reviews without a purchase: Low ratings, loyal customers, and deception. Journal of Marketing Research, 51(3), 249-269.

Ball, L., & Elworthy, J. (2014). Fake or real? The computational detection of online deceptive text. Journal of Marketing Analytics, 2(3), 187-201.

Banerjee, S., & Chua, A. Y. (2014). A theoretical framework to identify authentic online reviews. Online Information Review.

Banerjee, S., Bhattacharyya, S., & Bose, I. (2017). Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business. Decision Support Systems, 96, 17-26

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. Decision Support Systems, 50(2), 511-521.

Chen, L. S., & Lin, J. Y. (2013, July). A study on review manipulation classification using decision tree. In 2013 10th international conference on service systems and service management (pp. 680-685). IEEE.

Cheng, Y. H., & Ho, H. Y. (2015). Social influence's impact on reader perceptions of online reviews. Journal of Business Research, 68(4), 883-887.

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 1-24.

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and

SMOTE. Information Sciences, 465, 1-20.

Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., ... & Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. International Journal of Information Management, 59, 102168.

Eslami, S. P., & Ghasemaghaei, M. (2018). Effects of online review positiveness and review score inconsistency on sales: A comparison by product involvement. Journal of Retailing and Consumer Services, 45, 74-80.

Fernández, A., García, S., Luengo, J., Bernadó-Mansilla, E., & Herrera, F. (2010). Genetics-based machine learning for rule induction: state of the art, taxonomy, and comparative study. IEEE Transactions on Evolutionary Computation, 14(6), 913-941.

Filieri, R. (2015). What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. Journal of business research, 68(6), 1261-1270.

Gobi, N., & Rathinavelu, A. (2019). Analyzing cloud based reviews for product ranking using feature based clustering algorithm. Cluster Computing, 22(3), 6977-6984.

Gössling, S., Hall, C. M., & Andersson, A. C. (2018). The manager's dilemma: a conceptualization of online review manipulation strategies. Current Issues in Tourism, 21(5), 484-503.

He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. Marketing Science.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. Decision support systems, 52(3), 674-684.

Hu, N., Liu, L., & Sambamurthy, V. (2011). Fraud detection in online consumer reviews. Decision Support Systems, 50(3), 614-626.

Ismagilova, E., Slade, E., Rana, N. P., & Dwivedi, Y. K. (2020). The effect of characteristics of source credibility on consumer behaviour: A meta-analysis. Journal of Retailing and Consumer Services, 53, 101736.

Jalther, D., & Priya, G. (2019). Reputation reporting system using text based classification. Int. J. Innov. Technol. and Expl. Eng., 8(8), 1555-1558.

Khurshid, F., Zhu, Y., Xu, Z., Ahmad, M., & Ahmad, M. (2019). Enactment of ensemble learning for review spam detection on selected features. International Journal of Computational Intelligence Systems, 12(1), 387-394.

Kim, J., & Kwahk, K.-Y. (2022). Class Imbalance Resolution Method and Classification Algorithm Suggesting Based on Dataset Type Segmentation. Journal of Intelligence and Information Systems, 28(3), 23-43.

Kim, M. J., Kang, D. K., & Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. Expert Systems with Applications, 42(3), 1074-1082.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. GESTS international transactions on computer science and engineering, 30(1), 25-36.

Kumar, A., Gopal, R. D., Shankar, R., & Tan, K. H. (2022). Fraudulent review detection model focusing on emotional expressions and explicit aspects: investigating the potential of feature engineering. Decision Support Systems, 155, 113728.

Li, H., Li, J., Chang, P. C., & Sun, J. (2013).

Parametric prediction on default risk of Chinese listed tourism companies by using random oversampling, isomap, and locally linear embeddings on imbalanced samples. International Journal of Hospitality Management, 35, 141-151.

Li, L., Qin, B., Ren, W., & Liu, T. (2017). Document representation and feature combination for deceptive spam review detection. Neurocomputing, 254, 33-41.

Li, X., Yun, H., Li, Q., & Kim, J. (2022). A multi-channel CNN based online review helpfulness prediction model. Journal of Intelligence and Information Systems, 28(2), 171-189.

Liang, Y., & Zhu, K. (2018, April). Automatic generation of text descriptive comments for code blocks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

Lim, E. P., Nguyen, V. A., Jindal, N., Liu, B., & Lauw, H. W. (2010, October). Detecting product review spammers using rating behaviors. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 939-948).

Liu, Y., Pang, B., & Wang, X. (2019). Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph. Neurocomputing, 366, 276-283.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information sciences, 250, 113-141.

Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp. com. Com (March 15, 2016). Harvard Business School NOM Unit Working Paper, (12-016).

Majumdar, S., Kulkarni, D., & Ravishankar, C. V. (2007, May). Addressing click fraud in content delivery systems. In IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications (pp. 240-248). IEEE.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. American Economic Review, 104(8), 2421-55.

Mouratidis, D., Nikiforos, M. N., & Kermanidis, K. L. (2021). Deep learning for fake news detection in a pairwise textual input schema. Computation, 9(2), 20.

Nunamaker Jr, J. F., Burgoon, J. K., & Giboney, J. S. (2016). Information systems for deception detection. Journal of Management Information Systems, 33(2), 327-331.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557.

Park, Y.-J., & Kim, K.-j. (2017). Impact of Semantic Characteristics on Perceived Helpfulness of Online Reviews. Journal of Intelligence and Information Systems, 23(3), 29-44.

Rajamohana, S. P., & Umamaheswari, K. (2018). Hybrid approach of improved binary particle swarm optimization and shuffled frog leaping for feature selection. Computers & Electrical Engineering, 67, 497-508.

Rajamohana, S. P., Umamaheswari, K., & Abirami, B. (2017). Performance analysis of iBPSO and BFPA based feature selection techniques for improving classification accuracy in review spam detection. Appl. Math, 11(4), 1149-1153.

Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. Information Sciences, 385, 213-224.

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S. G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, 102771.

Scott, K. (2020). Microsoft teams up with OpenAI to exclusively license GPT-3 language model. Official Microsoft Blog.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2011). Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner. John Wiley and Sons.

Suh, Y., Yu, J., Mo, J., Song, L., & Kim, C. (2017). A comparison of oversampling methods on imbalanced topic classification of Korean news articles. Journal of Cognitive Science, 18(4), 391-437.

Tian, K., Shao, M., Wang, Y., Guan, J., & Zhou, S. (2016). Boosting compound-protein interaction prediction by deep learning. Methods, 110, 64-72.

Veganzones, D., & Séverin, E. (2018). An investigation of bankruptcy prediction in imbalanced datasets. Decision Support Systems, 112, 111-124.

Weisberg, J., Te'eni, D., & Arman, L. (2011). Past purchase and intention to purchase in e-commerce: The mediation of social presence and trust. Internet research.

Yelp Trust & Safety. Trust & Safety Report. https://trust.yelp.com/trust-and-safety-report/

Zhang, D., Zhou, L., Kehoe, J. L., & Kilic, I. Y. (2016). What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. Journal of Management Information Systems, 33(2), 456-481.

국문요약

기계학습과 GPT3를 시용한 조작된 리뷰의 탐지

체르냐예바 올가* · 홍태호**

고객의 구매 의사결정에 영향을 주는 온라인 리뷰의 부적절한 조작을 통해 이익을 얻고자 하는 기업 또는 온라인 판매자들 때문에, 리뷰의 신뢰성은 온라인 거래에서 매우 중요한 이슈가 되었다. 온라인 쇼핑몰 등에서 온라인 리뷰에 대한 소비자들의 의존도가 높아짐에 따라 많은 연구들이 조작된 리뷰를 탐지하는 방법에 개발하고자 하였다. 기존의 연구들은 온라인 리뷰를 기반으로 정상 리뷰와 조작된 리뷰를 대상으로 기계학습으로 이용함으로써 조작된 리뷰를 탐지하는 모형을 제시하였다. 기계학습은 데이터를 이용하여 이진분류 문제에서 탁월한 성능을 보여왔으나, 학습에 충분한 데이터를 확보할 수 있는 환경에서만 이러한 성능을 기대할 수 있었다. 조작된 리뷰는 학습용으로 사용할 수 있는 데이터가 충분하지 못하며, 이는 기계학습이 충분한 학습을 할 수 없다는 치명적 약점으로 내포하게 된다. 본 연구에서는 기계학습이 불균형 데이터 셋으로 인한 학습의 저하를 방지할 수 있는 방안으로 부족한 조작된 리뷰를 인공지능을 이용하여 생성하고 이를 기반으로 균형된 데이터 셋에서 기계학습을 학습하여 조작된 리뷰를 탐지하는 방안을 제시하였다. 파인 튜닝된 GPT-3는 초거대 인공지능으로 온라인 플랫폼의 리뷰를 생성하여 데이터 불균형 문제를 해결하는 오버샘플링 접근방법으로 사용되었다. GPT-3로 생성한 온라인 리뷰는 기존 리뷰를 기반으로 인공지능이 작성한 리뷰로써, 본 연구에서 사용된 로짓, 의사결정나무, 인공신경망의 성능을 개선시키는 것을 SMOTE와 단순 오버샘플링과 비교하여 실증분석을 통해서 확인하였다.

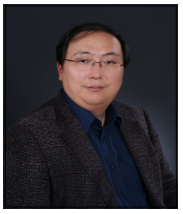
주제어 : 텍스트 마이닝, 온라인 리뷰, 조작된 리뷰 탐지, 텍스트 생성, 클래스 불균형 문제.

논문접수일 : 2022년 11월 20일　　논문수정일 : 2022년 12월 14일　　게재확정일 : 2022년 12월 15일
원고유형 : 일반논문　　교신저자 : 홍태호

* 부산대학교 경영학부
** 교신저자 : 홍태호
부산대학교 경영학부
부산광역시 금정구 부산대학로63번길 2 (장전동) 부산대학교 경영대학
Tel: 051-510-2531, E-mail: hongth@pusan.ac.kr

Biography

Olga Chernyaeva is a Ph.D. student of Management Information Systems at the College of Business Administration, Pusan National University. She received her Master's degree from Pusan National University. Her research interest includes business analytics, intelligent systems, data mining, and recommender systems for e-business. Her work has been published in the Asia Pacific Journal of Information Systems and the Journal of Information Systems.



Taeho Hong is a Professor of Management Information Systems at the College of Business Administration, Pusan National University. He received his Ph.D. from the Korea Advanced Institute of Science and Technology. His research interest includes intelligent systems, data mining, and recommender systems for e-business. His work has been published in Expert Systems with Application, Expert Systems, and Information Processing & Management.