

Machine Learning-Based Fake News Detection with Amalgamated Feature Extraction Method

Muhammad Bux Alvi^{1*}, Majdah Alvi¹, Rehan Ali Shah¹, Adnan Akhter¹, Mubashira Munir¹, Rakesh Kumar², Kavita Tabbassum³

Abstract:

Product fake reviews are increasing as the trend is changing toward online sales and purchases. Fake review detection is critical and challenging for both researchers and online retailers. As new techniques are introduced to catch the non-organic reviewer, so are their intruding approaches. In this paper, different features are amalgamated along with sentiment scores to design a model that checks the model performance under different classifiers. For this purpose, six supervised learning algorithms are utilized to build the fake review detection models, using LIWC, unigrams, and sentiment score features. Results show that the amalgamation of selected features is a better approach to counterfeit review detection, achieving an accuracy score of 88.76%, which is promising when compared to similar other work.

Keywords: *fake reviews, machine learning, amalgamated features, LIWC, sentiment score*

1 Introduction

A fake review is a false judgment or an opinionated text on a product or a service. Reviews can significantly affect the decision of buyers while shopping online. According to “Statista” statistics, e-commerce sales increase 6% in America from 2013 to 2020 [1]. As online purchase increases, so is the competition of online retailer giants. Therefore, the retailers and manufacturers take these reviews on a serious note. Fake reviewers capitalize on this opportunity to artificially devalue or promote products and services [2][3]. Hence, fake review prediction becomes a critical research area as online purchases increase. With the explosive growth of online businesses, the quantity and importance of reviews continue to increase. Fake reviews severely threaten researchers [4] and online retailers [5]. Reviews

can be positive to increase purchases on an online platform by manipulating users with fake customer reviews. Conversely, it can be a negative review to distract purchasers. It is estimated that 80% of users believe in posted product reviews before purchasing any product [6]. Negative fake reviews are used to defame competitor’s reputations. People who post such fake reviews are usually freelancers, and companies hire their services for writing fake reviews. Giant retailers like Amazon find these fake reviews of severe threat to their reputation and filed a complaint against review spamming [7].

Fake review prediction can be performed manually or automatically. Research has been carried out on manual opinion spam prediction for several years [8]. Early methods of fake review prediction were rudimentary. Many texts

¹Department of Computer Systems Engineering, Faculty of Engineering, The Islamia University of Bahawalpur, Pakistan

²Freelancer and Researcher

³Department Information Technology Center, Sindh Agriculture University Tandojam, Sindh, Pakistan

Corresponding Author: mbalvi@iub.edu.pk

analysis-based approaches are found in the literature [9]. Based on the research, commercial platforms developed opinion spam filtering systems to detect deceptive reviews.

Nevertheless, these systems make the fake reviewers enhance their review quality and deceive the detecting systems [10]. As time elapsed, those traditional approaches would not work efficiently because the fake reviewers started behaving like regular users.

Therefore, the trend of manual fake review prediction changed from text-based analysis to pattern and feature analysis like time [11], topics [12], ranking pattern [13], activity volume [14], and geolocation [15]. However, manual methods are slow, expensive, and of low accuracy. Automated methods based on machine learning could also identify the opinion spams and spammers by analyzing the review features. Text mining and Natural Language Processing (NLP) work together to generate the concept of content mining, and review spam detection comes under this concept. Additional review characteristics like review timings, reviewer id, and deviation trend of the review from other reviews of the same category are also considered in spam review detection. Jindal et al. [16] used the machine learning technique and showed that the amalgam of features is more robust than a single feature for fake review prediction. Li et al. in [17] showed that combining a bag of words (BOW) with more general features performs better than BOW alone. Mukherjee et al. [18] used machine learning with abnormal behavioral features of the reviewers and depicted that this technique was better than the linguistic features-based technique.

The significant contribution of this research is to develop a fake reviews detection model that uses machine learning techniques that will employ a heuristic optimization algorithm for affecting features and test its reliability and robustness against existing techniques. Such a model, when employed, can benefit retailers and giant business companies to shield their businesses against fake reviews and reviewers.

2 Literature Review

Advancements have been made in fake review detection by introducing new techniques and methods by researchers. These techniques play their role in improving accuracy and performance. So far, reviews are marked as spam based on either review spam detection or reviewer spam detection. Both techniques are helpful in fake review detection. Prior deals with content mining and natural language processing (NLP), whereas later technique applied on reviewer id and his behavior. Jindal et al. [16] is the first researcher who studied opinion spamming using supervised learning. The author divided the reviews into three categories (fake opinions, the brand only reviews, and non-review) and detected opinion spamming by finding duplicate reviews using the “w-shingling” method. The author used a dataset from Amazon with more than 5 million product reviews, applied his devised technique with a logistic regression algorithm, and achieved an AUC of 78%. Lim, Nguyen, Jindal, Liu, and Lauw [19] proposed a behavioral methodology for revealing spammers for review. They tried to figure out some spammer habits like targeting goods and tried to optimize their effect.

Moreover, they suggested a model focused on specific patterns to identify rating spammers. Ott et al. [20][21] created a data set for analysis in review spam detection. The data set comprised positive opinion spam with truthful reviews and negative opinion spam with real reviews. The author applied the n-gram and linguistic features to find fake reviews under a supervised learning mechanism, and the results were verified with human performance. In their research, Feng et al. [4] framed a model based on the normal distribution of opinion to detect fake reviews. In their view, a product or a service review involved this concept of normal distribution of opinion. Shojaee et al. [9] suggested a novel technique for fake review detection by combining Lexical and Synthetic features. Elmurngi and Gherbi [22] proposed a text classification and sentimental analysis approach for different machine learning algorithms with stop words and without stop words. They also applied a

decision tree algorithm to improve their results. Shah, Ahsan, Kafi, Nahian, and Hossain [23] combined Supervised & Active learning and created a model to detect spamming. Both fictitious and real-life data were used for spam analysis.

3 Proposed Approach

This section describes the proposed method to accomplish the task of fake review detection. This research uses two features as classification criteria with a sentiment score feature (an additional feature). These individual features and combinations are used to train various classifiers and tested against evaluation metrics. Reviews are classified as fake or not fake. This study uses six classification algorithms: Naive Bayes, decision tree, instance-based KNN, support vector machine (SVM), logistic Regression, and Random Forest. Training data is 80%, and 20% of data is set aside for testing purposes with a 5-fold cross-validation technique. Figure 1 presents the adopted research method for this work.

3.1 Data Acquisition and Pre-processing

The data set selected for this research contains 1600 reviews combined from two data sets (hotel review data sets). The data sets were created by Myle et al. and are available from [20][21]. The data set contains eight hundred truthful reviews, of which four hundred are positive, and four hundred are negative. Similarly, 800 spam reviews are also included in this data set, of which half are positive, and half are negative. The preprocessing of the data set significantly affects the accuracy of results [24][25][26]. Furthermore, preprocessing curbs feature vector space. Therefore, preprocessing techniques like missing values management, tokenization, stop words removal, and generating n-gram are implemented on the data set to obtain cleaner data set.

3.2 Features

Features are pieces (s) of text that have semantic significance. In the text data systems,

features highly influence the effectiveness of the developed model.

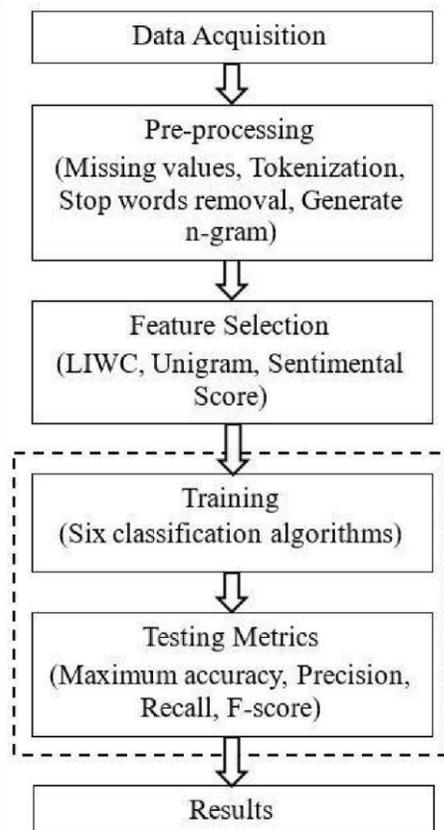


Fig. 1. Proposed Machine Learning Approach with Amalgamated Features for Fake Reviews Detection

3.2.1 N-Grams

In this feature extraction method, n-adjacent tokens are picked as a feature from review contents. It is denoted as unigram if one adjacent word is selected, bigram if two adjacent words are selected, and trigram with three adjacent words at a time. These features can effectively help model all the content within the text. In this research work, unigram is used as a feature.

TABLE I. STATE OF ART SPAM REVIEW DETECTION TECHNIQUES

Reference	Year	Data set	Learning type	Techniques/ Algorithm	Results	Limitations
[4] Feng	2012	Ott et al. data set with modification	Supervised Learning	LIBSVM classifier/ Term frequency	Accuracy 72.5%	Specific kind of dataset
[9] Shojaei	2013	Ott et al. data set	Supervised Learning	SVM/ Naïve Bayes/ Stylometric Feature	F-measure 84%	Limited to a specific domain
[13] Jindal N, Liu B	2007	Data set of the manufactured product only	Supervised Learning	Logistic Regression	average AUC 78%	Lack of accuracy of a real-world data set
[18] Lim, Ee-Peng	2010	Amazon Data set	Supervised Learning	Behavioral features of Spammer	Accuracy 78%	Limit set of data for supervised learning
[20] Jeffrey T. et al	2013	Ott et al. data set	Supervised Learning	Support Vector Machine (SVM)	Accuracy 86 %	Human judgments can be imperfect and biased.
[21] Elmurngi E.	2017	Movie review data set	Supervised Learning	DT(DT-J48)/ SVM/KNN	Accuracy 81.75%	Feature selection methods are not used
[22] Ahsan, Nahian, Kafi, Hossain and Shah	2017	Ott dataset	Active/ Supervised Learning	Hybrid classifier using NB/ SVC /DT /Maximum Entropy	Accuracy 95%	Small scale dataset is used for a specific domain

3.2.2 LIWC

The Linguistic Inquiry and Word Count (LIWC) is a text analysis method. This method can analyze eighty different features, for example, psychological concerns like emotion, text functional aspects, and personal and perception concerns like religion [27].

3.2.3 Sentiment Score

It has been observed that spammers with negative reviews generally use more negative words like “bad” and “dissatisfied”. This way,

the degree of negative sentiment is increased compared to a non-spam negative review. Likewise, spammers with positive reviews generally use positive terms such as “good”, “great”, “nice”, and “gorgeous”. Therefore, reviewers show more positive sentiment than a non-spam positive review. The sentiment score of a review can be calculated by the following formula [28].

$$SC(rt) = \sum (-1)^n \frac{S(W_i)}{\text{Distance}(fet_i, W_i)} \quad (1)$$

where "rt" is review text, "S(W_i)" is the sentiment polarity of word W_i (+1 or -1), "n" denotes the total number of negation-words in a feature with default = 0, "fet" refers to a feature in a review sentence and "distance (fet, W_i)" is the distance between feature and word.

3.3 Classification Algorithms

Six various classification algorithms are used in this paper in order to determine the effect of different features and their combinations on classification accuracy and performance.

3.3.1 Naïve Bayes (NB)

NB is based on the Bayes theorem [29]. It is a probabilistic multiclass classification algorithm assuming features independency to foresee the output class. Equation 2 checks the probability of the feature-set being categorized into a particular class:

$$P(x) = P(x_1)P(x_2)P(x_3)P(x_n) \quad (2)$$

where " $x = (x_1, x_2, \dots, x_n)$ " are a set of features. Individual probabilistic classification of a feature may be calculated as given in equation 3:

$$P(x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (3)$$

3.3.2 Decision Tree (DT)

The working principle of DT is based on a hierarchical breakdown of the data set used for training. In this classifier, features are used for labeling tree nodes, and the branches between them are given the weight representing the occurrence of feature in the test data; finally, class names are assigned to the leaf. The data set is divided into the presence or absence of features. The data set is divided recursively until the leaf nodes are reached.

Entropy Formula:

$$\text{Entropy} = - \sum_{j=1}^m P_{ij} \log_2 P_{ij} \quad (4)$$

3.3.3 Random Forest (RF)

RF is a voting method where many decision trees are grown simultaneously. The input features are fed to individual trees in the forest. The final classification is based on the overall most votes from all trees in the forest [30]. The mathematical form of random forest to calculate mean square error is:

$$f^{\wedge} = \sum_{s=1}^S \frac{1}{S} (fs - ys)^2 \quad (5)$$

Where "S" denotes the number of data points, "fs" is the value returned by the model, and "ys" is the actual value of data points.

3.3.4 Support Vector Machine (SVM)

SVM is a classification algorithm that finds the maximum margin hyperplane to classify the " i^{th} " vector. Optimal " y_i " (y_i denotes the target), " X_i " hyperplane is found by linear features between two classes (0 or 1).

3.3.5 K-nearest neighbor (KNN)

KNN is an instance-based algorithm that assumes that similar things exist in close proximity. In this technique, the feature is classified by the plurality vote of its neighbors by calculating their distances. It uses Euclidean distance formula to compute the distance between the points, which is mathematically represented as:

$$D = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \quad (6)$$

3.3.6 Logistic Regression

Logistic Regression is a model-based algorithm often used when the dependent variable is dichotomous in nature. However, it can be tuned to be used with multiclass classification tasks as well. Logistic Regression describes

the data set and defines the relationship between one dependent binary variable with one or more independent variables.

3.4 Testing Metrics

Accuracy, precision, recall, and f-score are used to evaluate model performance. These metrics can be defined as:

$$Accuracy = \frac{TuP + TuN}{TuP + TuN + FaP + FaN} \quad (7)$$

$$Precision = \frac{TuP}{TuP + FaP} \quad (8)$$

$$Recall = \frac{TuP}{TuP + FaN} \quad (9)$$

$$F - Score = \frac{2 * (recall * precision)}{(recall + precision)} \quad (10)$$

where "TuN", "TuP", "FaN", and "FaP" are true negative, true positive, false negative, and false positive respectively.

4 Experimental Results, Discussion and Evaluation

This section describes the experimental results, discusses the results, and evaluates the developed model quantitatively. Six machine learning algorithms (Naïve Bayes, decision tree, Random Forest, SVM, K-nearest neighbor, and Logistic Regression) were used to develop the model using three feature extraction techniques (LIWC, n-gram (unigrams), and sentiment score).

The results of individual feature and their combinations are shown in table 2. An accuracy of 63.22% is achieved when LIWC is used alone, but when combined with a sentiment score, accuracy increases to 70.35%. Classification model using unigram feature alone gives an accuracy of 73.55%, but if combined with sentiment score, it increases to

80.34%. Maximum accuracy of 88.76% is attained by combining LIWC, unigram, and sentiment scores. Eventually, this study supports and proves the initial hypothesis of getting improved results by using amalgamated features with machine learning algorithm-based classification models.

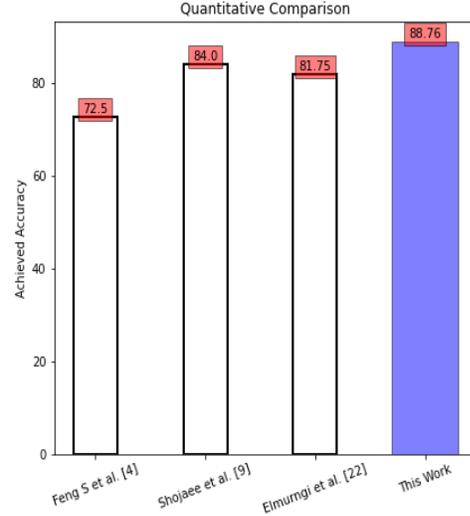


Fig. 2. Quantitative Comparison with Previous Similar Work

4.1 Qualitative Evaluation

To strengthen the hypothesis, Figure 2 shows the quantitative comparison of previous work [4][9][22] and the proposed method in this work to detect fake reviews using the Hotel reviews dataset. It indicates that the undertaken work supersedes the other work. Since a balanced dataset is used, the accuracy score measure is good enough for quantitative performance comparison of the developed machine learning algorithm-based models with previous work. This improvement is, at least, 4-5 points when compared with previous similar work.

5 Conclusion

This work was an effort to determine the effective combination of features that performs well for fake reviews detection. The study work used n-gram, LIWC, and sentiment

score features for training purpose. Different classifiers are trained on these features. The classification algorithms we chose are described in section 3.3. After experimental work, logistic Regression outperformed all other machine learning algorithm-based models. As far as features performance is concerned, unigram proved to be better when applied in separation than LIWC. However, the combination of both (unigram + LIWC) with sentiment score performed more adequately,

giving maximum accuracy of 88.76%. This result is better than some other techniques described in section 2. For future work, it is suggested to use semi-supervised learning to check the accuracy and performance of unigram and LIWC features on the fake review detection method. In this way, possible performance enhancement will be measured. At the same time, the limitation of the labeled data set for supervised learning will also be resolved.

TABLE II. PERFORMANCE EVALUATION OF DIFFERENT FEATURE EXTRACTION METHODS

Approaches (Features)	Maximum accuracy (%)	Precision	Recall	F-score
LIWC	63.22	58.00	64.43	61.05
Sentiment score, LIWC	70.35	62.50	73.88	67.71
Unigram	73.55	78.00	74.50	76.21
Sentiment score, Unigram	80.34	91.50	77.59	83.97
Sentiment score, LIWC, Unigram	88.76	92.00	82.61	87.05

REFERENCES

- [1] “U.S. e-commerce share of retail sales 2021-2025 | Statista.” <https://www.statista.com/statistics/379112/e-commerce-share-of-retail-sales-in-us/>.
- [2] F. Li, M. Huang, Y. Yang, and X. Zhu, “Learning to identify review spam,” *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 2488–2493, 2011, doi: 10.5591/978-1-57735-516-8/IJCAI11-414.
- [3] R. Y. K. Lau, S. Y. Liao, R. Chi-Wai Kwok, K. Xu, Y. Xia, and Y. Li, “Text mining and probabilistic language modeling for online review spam detection,” *ACM Trans. Manag. Inf. Syst.*, vol. 2, no. 4, Dec. 2011, doi: 10.1145/2070710.2070716.
- [4] S. Feng, L. Xing, A. Gogar, and Y. Choi, “Distributional footprints of deceptive product reviews,” *ICWSM 2012 - Proc. 6th Int. AAAI Conf. Weblogs Soc. Media*, pp. 98–105, 2012.
- [5] Sussin, J., and E. Thompson, “The consequences of fake fans, ‘Likes’ and reviews on social networks,” *Gart. Res.*, vol. 2091515, 2012.
- [6] “Amazon sues to block alleged fake reviews on its website | Reuters.” <https://www.reuters.com/article/us-amazon-com-lawsuit-fake-reviews-idUSKBN0N02LP20150410>.
- [7] “Local Consumer Review Survey 2022: Customer Reviews and Behavior.” <https://www.brightlocal.com/research/local-consumer-review-survey/>.
- [8] N. Spirin and J. Han, “Survey on Web Spam Detection: Principles and Algorithms,” vol. 13, no. 2, pp. 50–64.
- [9] N. M. S. and S. N. Somayeh Shojaee, Masrah Azrifah Azmi Muradt, Azreen Bin Azman, “Detecting Deceptive Reviews Using Lexical and Syntactic Features,” pp. 219–223, 2013.
- [10] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, “Automated crowdturfing attacks and defenses in online review systems,” *Proc. ACM Conf. Comput. Commun. Secur.*, pp. 1143–1158, 2017, doi: 10.1145/3133956.3133990.

- [11] K. C. Santosh and A. Mukherjee, "On the temporal dynamics of opinion spamming: Case studies on yelp," 25th Int. World Wide Web Conf. WWW 2016, pp. 369–379, 2016, doi: 10.1145/2872427.2883087.
- [12] S. Nilizadeh et al., "Poised: Spotting twitter spam off the beaten paths," dl.acm.org, pp. 1159–1174, Oct. 2017, doi: 10.1145/3133956.3134055.
- [13] H. Chen, "Toward Detecting Collusive Ranking Manipulation Attackers in Mobile App Markets," pp. 58–70, 2017.
- [14] D. Y. T. Chino, A. F. Costa, A. J. M. Traina, and C. Faloutsos, "VOLTIME: Unsupervised anomaly detection on users' online activity volume," Proc. 17th SIAM Int. Conf. Data Mining, SDM 2017, pp. 108–116, 2017, doi: 10.1137/1.9781611974973.13.
- [15] R. Deng, N. Ruan, R. Jin, Y. Lu, and W. Jia, "SpamTracer: Manual Fake Review Detection for O2O Commercial Platforms by Using Geolocation Features," pp. 1–20.
- [16] N. Jindal and B. Liu, "Review spam detection," 16th Int. World Wide Web Conf. WWW2007, pp. 1189–1190, 2007, doi: 10.1145/1242572.1242759.
- [17] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a General Rule for Identifying Deceptive Opinion Spam," pp. 1566–1576, 2014.
- [18] A. Mukherjee, V. Venkataraman, ... B. L.-S. international A., and U. 2013, "What Yelp fake review filter might be doing?," Proc. Int. AAAI Conf. Web Soc. Media, vol. 7, no. 1, pp. 409–418, 2011.
- [19] B. Liu and H. W. Lauw, "Detecting Product Review Spammers using Rating Behaviors," pp. 939–948, 2010.
- [20] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," ACL-HLT 2011 - Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol., vol. 1, pp. 309–319, 2011.
- [21] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," NAACL HLT 2013 - 2013 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Main Conf., no. June, pp. 497–501, 2013.
- [22] E. Elmurghi and A. Gherbi, "An empirical study on detecting fake reviews using machine learning techniques," 7th Int. Conf. Innov. Comput. Technol. INTECH 2017, no. Intech, pp. 107–114, 2017, doi: 10.1109/INTECH.2017.8102442.
- [23] M. N. I. Ahsan, T. Nahian, A. A. Kafi, M. I. Hossain, and F. M. Shah, "An ensemble approach to detect review spam using hybrid machine learning technique," 19th Int. Conf. Comput. Inf. Technol. ICCIT 2016, pp. 388–394, 2017, doi: 10.1109/ICCITECHN.2016.7860229.
- [24] W. Etaiwi and G. Naymat, "The Impact of applying Different Preprocessing Steps on Review Spam Detection," Procedia Comput. Sci., vol. 113, pp. 273–279, 2017, doi: 10.1016/j.procs.2017.08.368.
- [25] M. B. Alvi, N. A. Mahoto, M. A. Unar, and M. A. Shaikh, "An Effective Framework for Tweet Level Sentiment Classification using Recursive Text Preprocessing Approach," no. July, 2019, doi: 10.14569/IJACSA.2019.0100674.
- [26] M. B. Alvi, N. A. Mahoto, M. Alvi, M. A. Unar, and M. Akram Shaikh, "Hybrid classification model for twitter data-A recursive preprocessing approach," 5th Int. Multi-Topic ICT Conf. Technol. Futur. Gener. IMTIC 2018 - Proc., 2018, doi: 10.1109/IMTIC.2018.8467221.
- [27] C. G. Harris, "Detecting deceptive opinion spam using human computation," AAAI Work. - Tech. Rep., vol. WS-12-08, pp. 87–93, 2012.
- [28] P. Cavallo et al., "Journal of Software," vol. 9, no. 8, 2018.
- [29] M. Ben-bassat, K. L. Klove, and M. A. X. H. Weil, "CALO ($x = ALO(x)e$)," vol. 2, no. 3, pp. 261–266, 1980.
- [30] A. Akhter, M. B. Alvi, and M. Alvi, "Forecasting Multan estate prices using optimized regression techniques," Univ. Sindh J. Inf. Commun. Technol., vol. 5, no. 4 SE-Computer Science, Apr. 2022, <https://sujo.usindh.edu.pk/index.php/USJICT/article/view/4340>.
- [31] G. Mujtaba and E. S. Ryu, "Client-Driven Personalized Trailer Framework Using Thumbnail Containers," IEEE Access, vol. 8, pp. 60417–60427, 2020, doi: 10.1109/ACCESS.2020.2982992