Fake Online Reviews Detection and Analysis Using Bert Model

Dr. Chandaka Babi[1], M. Sai Roshini[2], P. Manoj[3], K. Satish Kumar[4]

[1]Assoc. Prof. in Department of Computer Science and Engineering at Raghu Engineering College, Visakhapatnam,India.

[2] Student in Department of Computer Science and Engineering at Raghu Engineering College, Visakhapatnam, India.

[3] Student in Department of Computer Science and Engineering at Raghu Engineering College, Visakhapatnam, India.

[4] Student Department of Computer Science and Engineering at Raghu Engineering College, Visakhapatnam, India.

E-mail address: babi.chandaka@raghuenggcollege.com , 19981a0599@raghuenggcollege.com, 19981a05c3@raghuenggcollege.com , 19981a0568@raghuenggcollege.com

Abstract

The impact of user reviews on the revenue of an e-commerce organization cannot be overstated. Consumers heavily depend on online reviews while taking decisions on purchasing, making credibility of these reviews crucial to businesses. Unfortunately, some companies resort to pay some people to post deceiving reviews, which can mislead consumers and harm a company's reputation. While various techniques have been developed to detect fake reviews over the past decade, there is a lack of comprehensive surveys that analyze and summarize these approaches. This paper aims to bridge the gap by focusing on detecting the fake reviews. It provides an overview of the current datasets available and their collection methods, as well as an analysis of the feature extraction techniques utilized in prior research. We employ statistical ML techniques, including Naïve Bayes, Support Vector Machines, and the transformer BERT, to conduct experiments on Twitter review datasets. Our results show that the Naïve Bayes algorithm achieved an accuracy of 97.14%, while Support Vector Machines achieved an accuracy of 100%. These results provide a baseline for future studies in this area.

Keywords: Fake Reviews, Statistical Machine Learning Techniques, Supervised Models, BERT, Naïve Bayes, E-commerce, Support Vector Machine, Twitter dataset.

1. INTRODUCTION

In today's digital age, customers have the ability to post their reviews and opinions on a variety of websites. Online reviews have become a crucial factor in the decision-making process of both consumers and businesses[6]. They play an important role in shaping product design and marketing strategies, while also providing valuable information for potential buyers[7]. As a result, the number of customer reviews has significantly increased over time, and they pose significant effect on the purchasing decisions of potential customers[9]. Positive reviews can bring significant financial benefits to businesses, while negative reviews can have a detrimental effect on their reputation[9]. Therefore, as the demand for customer feedback continues to grow, the importance of incorporating it into business strategies to improve products, services, and marketing becomes increasingly evident.

Social media platforms provide an open platform for individuals to share their opinions and reviews about any business at any time, without any restrictions or obligations[11]. This lack of regulations can lead some businesses to misuse social media platforms for the unfair promotion of their products, brands, or stores, and to unfairly criticize their competitors. For instance, if a group of customers who bought a particular laptop post negative reviews about its display, gpu performance, it may create a negative perception of all laptops among the public[14]. In response, the laptop manufacturers may hire people or groups to post false positive reviews of their products, while also hiring staff to post negative reviews of their competitors' products, to promote their own business.

It is challenging to attain high levels of accuracy in identifying fake reviews due to the prevailing state of such reviews. Despite attempts to utilize vast datasets to train machine learning models, detecting fake reviews still poses a significant difficulty.

Fake reviews can have a significant impact on the reputation and finances of many businesses. To address this issue, we are implementing machine learning methods to detect fake reviews. Our project utilizes SVM, NB, and BERT transformer models to identify and flag potentially fraudulent reviews.

The authors confirm contribution to the paper as follows: M. Sai Roshini and P. Manoj conceived and designed the study, while P. Manoj collected the data. M. Sai Roshini, P. Manoj, and K. Satish Kumar analyzed, interpreted the results, and drafted the manuscript. All authors participated in reviewing the results and approving the final version of the manuscript.

This paper consists of the following sections: Literature Review, Existing Work, Methodology consisting of the Proposed System, containing Data collection, Data preprocessing, Model training, Algorithms used, Model Testing and Evaluation; Results and Analysis, Conclusion and Future Enhancements followed by References.

2. LITERATURE REVIEW

The issue of identifying false reviews, especially in the form of review spam, has been the subject of thorough investigation since 2007. In a prior research, the authors examined Amazon and discovered that identifying fake reviews through manual reporting can be challenging. This is because fake reviews are often crafted to appear credible to other users. To address this challenge, the authors proposed a model to detect fake reviews using duplicate or nearly identical reviews as spam

Previous studies on fake review detection focus primarily on textual and behavioural characteristics, although some studies also consider social or temporal factors. Several articles suggest using text characteristics to identify fake reviews.

E. P. Lim et al., aimed to identify and track comment spammers by modelling their characteristic behaviour. The authors propose a scoring method to measure each reviewer's spam level and apply it to the Amazon review data set. Their results show that their proposed ranking and monitoring method is effective in spotting spammers and outperforms other benchmark methods

based solely on voting support [1].

Ott. Myle et al., studies fraudulent opinion spam - fictitious opinions deliberately written to ring true. The researchers created and evaluated three techniques for identifying deceptive opinion spam and finally constructed a classifier that achieves an accuracy rate of almost 90% on their evaluation dataset for opinion spam. Their work sheds light on the relationship between misleading perspectives and imaginative writing [2].

J. K. Rout et al., proposed: online opinion reviews play a crucial role in informing consumers' decision-making processes, which can have a huge impact on businesses' profitability. Regrettably, there has been an escalation in individuals or organizations who take advantage of online reviews for their personal benefit, which has resulted in a surge in misleading and counterfeit reviews. This has sparked interest in research on detecting such reviews. [3]

3. EXISTING WORK

The existing method has primarily concentrated on various supervised machine learning classification algorithms. These algorithms aim to establish an appropriate model for distributing the training data. Techniques such as SVM, NB, DT, linear SVC have been employed to train the data and achieved an accuracy of up to 90%[12]. However, these techniques were insufficient in effectively training extensive datasets. The proposed system aims to address these shortcomings.

4. METHODOLOGY

4.1. PROPOSED SYSTEM

The goal of our work is to get higher accuracy levels, training on large datasets and creating a user-friendly webpage for fake review detection. So, we proposed a BERT based Machine Learning approach where it performs feature extraction. BERT is most suitable method for semantic based reviews. The proposed work is shown in Figure 1.

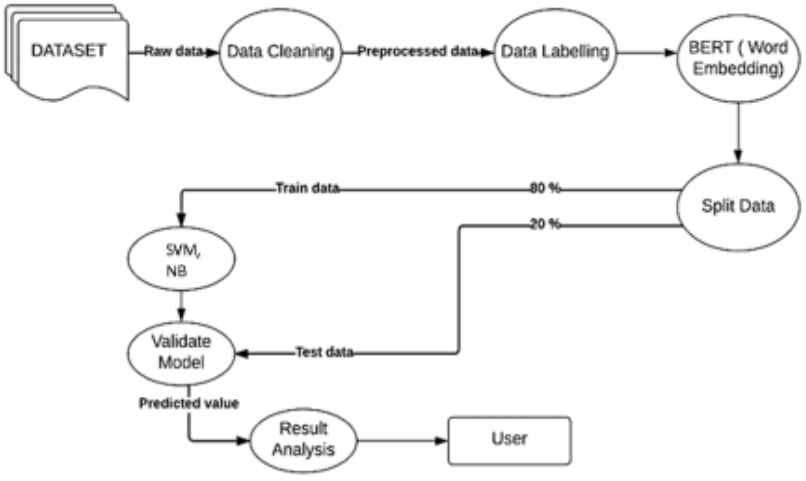

Figure 1. Flowchart

The subsequent section outlines the components integrated into the proposed system.

a. Data Collection

We utilized a Twitter dataset which was previously labeled to detect fake online reviews. The dataset contains 1600 online reviews and is categorized into 5 categories: Hotel, Polarity, Source, Result, and Text. Through the use of this dataset, we were able to achieve high levels of accuracy in our detection of fake reviews. The dataset considered is from the figure 2 and figure 3.

	A	B	C	D	E
1	deceptive	hotel	polarity	source	text
2	truthful	conrad	positive	TripAdvisor	We stayed for a one night getaway with family on a thursday. Triple AAA rate of 173
3	truthful	hyatt	positive	TripAdvisor	Triple A rate with upgrade to view room was less than $200 which also included
4	truthful	hyatt	positive	TripAdvisor	This comes a little late as I'm finally catching up on my reviews from the past several
5	truthful	omni	positive	TripAdvisor	The Omni Chicago really delivers on all fronts, from the spaciousness of the rooms to
6	truthful	hyatt	positive	TripAdvisor	I asked for a high floor away from the elevator and that is what I got. The room was
7	truthful	omni	positive	TripAdvisor	I stayed at the Omni for one night following a business meeting at another downtown
8	truthful	conrad	positive	TripAdvisor	We stayed in the Conrad for 4 nights just before Thanksgiving. We had a corner room
9	truthful	omni	positive	TripAdvisor	Just got back from 2 days up in Chicago shopping with girlfriends. First time I have
10	truthful	omni	positive	TripAdvisor	We arrived at the Omni on 2nd September for a 6 day stay. I took ill when I left the
11	truthful	hyatt	positive	TripAdvisor	On our visit to Chicago, we chose the Hyatt due to its location in downtown, within
12	truthful	fairmont	positive	TripAdvisor	I stayed at the Fairmont Chicago for one night - I'm a frequent business traveler and
13	truthful	conrad	positive	TripAdvisor	Ok, so first trip to chicago and I was a litlle worried about the hotel and the location,
14	truthful	hyatt	positive	TripAdvisor	We arrived at 10:30 am on a Friday, and they had a room ready for us by 11:30 am,
15	truthful	conrad	positive	TripAdvisor	My wife and I came to spend the weekend in downtown Chicago for shopping and we

Figure 2. The dataset contains real reviews.

	A	B	C	D	E
407	deceptive	hotel	polarity	source	text
408	deceptive	conrad	positive	MTurk	My visit to the Conrad Chicago in the heart of downtown was a luxurious experience
409	deceptive	omni	positive	MTurk	I recently had my wedding at the Omni Hotel in Chicago and I was more than pleased.
410	deceptive	conrad	positive	MTurk	If you are looking for a luxurious downtown Chicago experience, the Conrad Chicago is
411	deceptive	conrad	positive	MTurk	My wife and I checked in to this hotel after a rough flight from Los Angeles. We
412	deceptive	conrad	positive	MTurk	We stayed here for a weekend trip to Chicago and we will come back and stay here
413	deceptive	fairmont	positive	MTurk	WOW! That is all I can say! My girlfriends and I recently went on a little 'excursion' to
414	deceptive	omni	positive	MTurk	Luxury and comfort combine to give a wonderful sense of relaxation for guests at the
415	deceptive	hyatt	positive	MTurk	Recently, I traveled to Chicago for a business conference and was lucky enough to stay
416	deceptive	conrad	positive	MTurk	I had a great experience staying at the Conrad Chicago, the service was top notch, the
417	deceptive	hyatt	positive	MTurk	My husband and I stayed at the Hyatt Regency while attending a family wedding in
418	deceptive	hyatt	positive	MTurk	From the moment I stepped into the hotel I felt like I was walking into my own castle
419	deceptive	omni	positive	MTurk	We stayed at the Omni Chicago Hotel recently and had a great experience! My
420	deceptive	fairmont	positive	MTurk	The Fairmont Chicago Millenium Park Hotel provides a rich, beautiful, cultured
421	deceptive	hyatt	positive	MTurk	My husband and I stayed at the Hyatt Regency Chicago to celebrate our ten year
422	deceptive	hyatt	positive	MTurk	Overall, the hotel is beautiful. The service was great. All the employees were friendly

Figure 3. The dataset contains fake reviews.

b. Data Preprocessing

For data pre-processing, we combined the text data, created a dictionary, and assigned numerical values to the text. In our research, we opted to use word frequency count, sentiment polarity, and review length as features for our analysis.

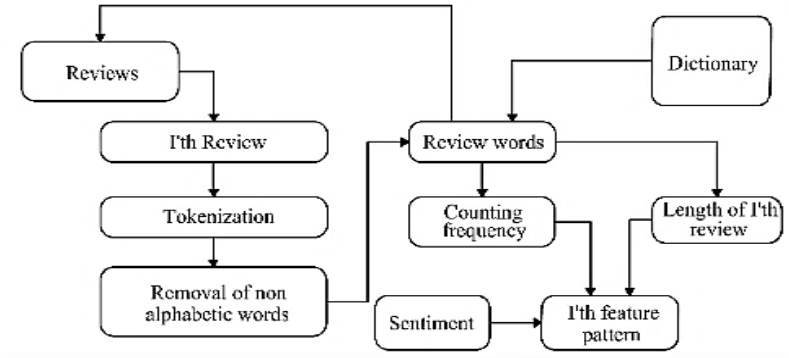

Figure 4. Feature extraction diagram.

c. Model Training

The process of model training is essential after building the model. Data preprocessing is a crucial step that affects the accuracy of the model. Features are selected to improve the accuracy of the model. This involves manually extracting frequency and time domain features. We will apply SVM and NB algorithms on BERT vector to train the algorithms.

d. Algorithms used:

1. Naïve Bayes:

Naive Bayes is widely used for mining the text, as it can be applied wherever the conditional independence is maintained. This property is particularly useful for analyzing text, where the sequence and content of words can vary greatly. Being a probabilistic method, the Naive Bayes classifier can be used for both classification and regression and is known for its speedy calculation.

2. Support Vector Machine:

Support vector machine is a ML algorithm which creates a hyperplane or multiple hyperplanes in high-dimensional space to classify or perform regression tasks. The SVM classifier's parameters were fine-tuned to improve our results. SVM is suitable for modifying high-dimensional datasets and can be used to classify fake review detection data in various fields, including biological data.

3. BERT:

Bidirectional Encoder Representations from Transformers is a pre-trained deep learning model which has been fine-tuned for several natural language processing tasks, like sentiment analysis and language translation. BERT can understand the context of a sentence or phrase, particularly important for detecting fake reviews. It leverages the contextual understanding and accurately determine whether a review is genuine or fake.

e. Model Testing and Evaluation

After training the model, it is important to test it on a separate dataset to assess its generalization ability. The testing dataset should be distinct from the training dataset and should be labeled with the true class labels of the reviews. Using SVM and NB classifiers we complete the training. The dataset results are tested by using 30% of it. The predicted labels are compared with the true labels to calculate the model's accuracy.

- The effectiveness of models is measured by evaluating the accuracy, where it refers to the proportion of correctly classified samples to all possible samples.

5. RESULTS AND ANALYSIS

For feature extraction, we have employed the BERT algorithm. As for the classifiers, we have utilized the Naïve Bayes and Support Vector Machine classifiers. To conduct each classification process, we partitioned the dataset into the train-test ratio 70:30.

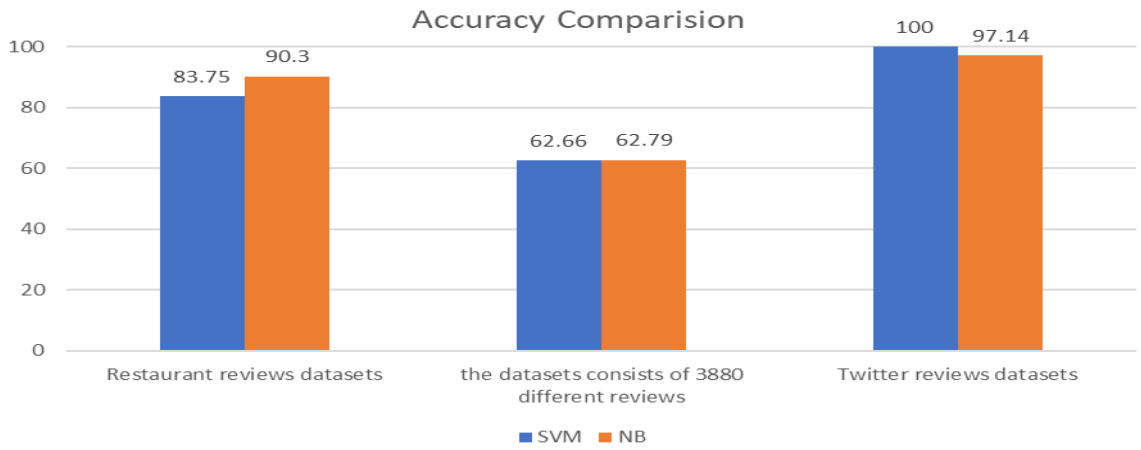

Figure 5. The accuracy comparison graph

From Figure 5, we can see the Support Vector Machine classifier has got an accuracy of 100% whereas Naïve Bayes classifier has got an accuracy of 97.14 % with a ratio of 70:30 respectively.

In previous publications, Author Chapalamadugu Haritha Chowdary used Naïve Bayes, Decision Tree and Support Vector Machine ML algorithms on Restaurant review datasets. So, they obtained an accuracy of 90.3% for NB and 83.75% for SVM.

Author Dogo Rangsang using Semi-Supervised classification with Expectation Maximization algorithm containing SVM and NB on the dataset which contains 3880 distinct reviews achieved the accuracy where 62.79 % for NB and 62.66 % for SVM.

In figure 6, the start screen of the Fake Online Review Detection is displayed which consists of the title of the application, a text box followed by a Submit Button This is the start screen where it prompts the user to enter the review which they wish to verify. In Figure 7, review 1 is entered which needs to be checked if it is a Genuine or a Fake review. And then the review is submitted. In figure 8, the result screen is displayed which shows that the “review is not fake”. Figure 9, represents review two which needs to be checked. In Figure 10, the result is displayed that the “review is fake”.

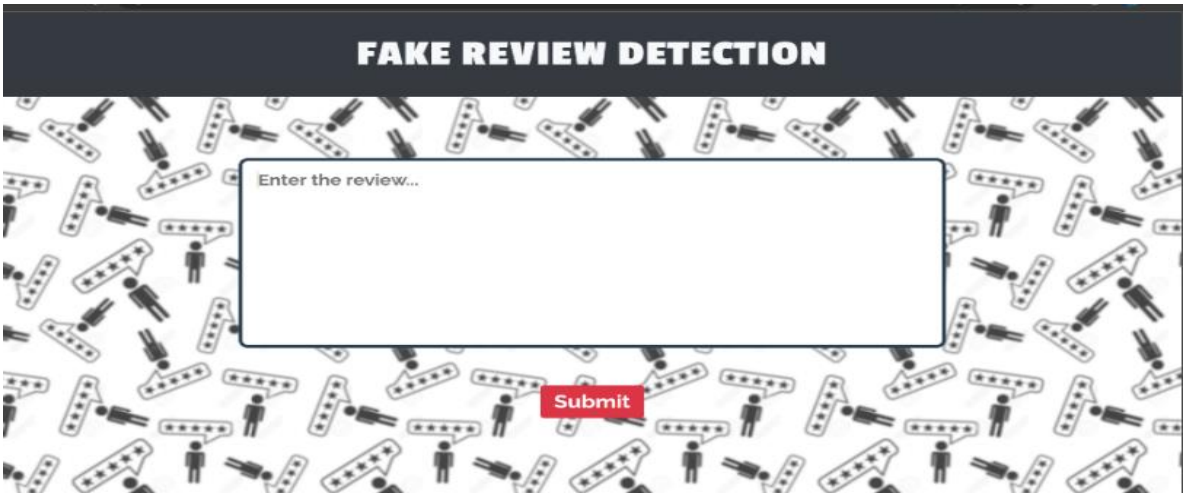

Figure 6. The Start-screen

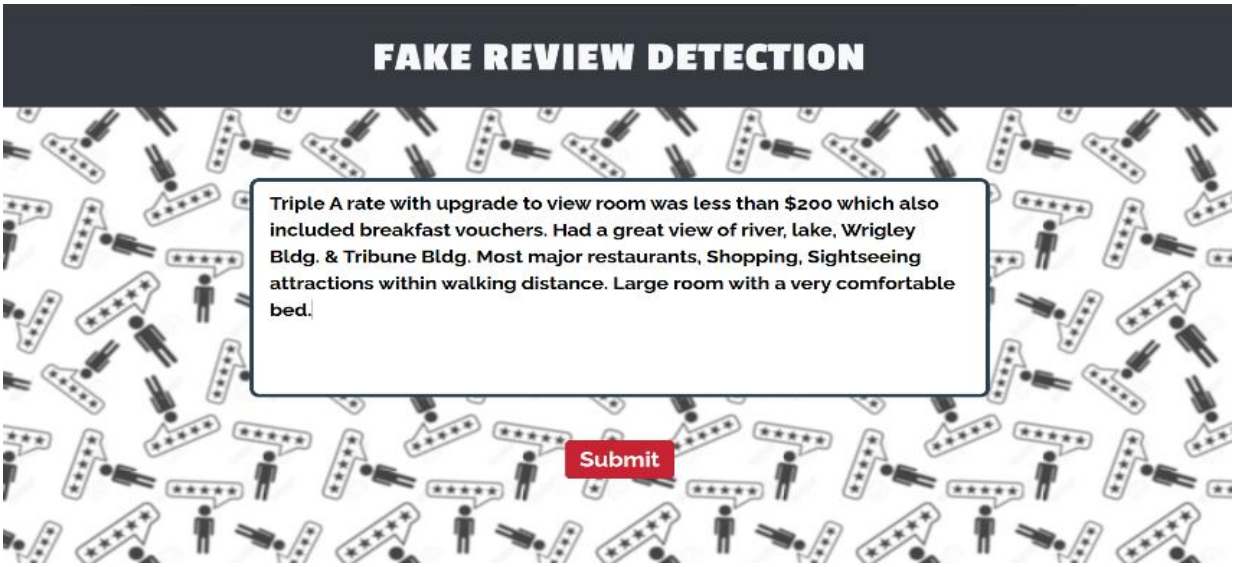

Figure 7. Review one

Figure 8. The output for Review one

Figure 9. Review two

Figure 10. The output for Review two

6. CONCLUSION AND FUTURE ENHANCEMENTS

This review paper provides a detailed analysis of machine learning-based techniques for detecting fake reviews. Firstly, it examines the feature extraction methods that have been used by several researchers. Secondly, it discusses the construction of methods and pre-processed datasets. The review notes that statistical ML techniques can improve text classification performance by enhancing classifier design and feature extraction. Our project used BERT to obtain the highest accuracy of 97.14% for Naïve Bayes classifier and 100% for Support Vector Machines.

Furthermore, the review paper summarizes current research gaps and proposes potential future directions for achieving robust outcomes. Most of the current works in this field concentrate on supervised machine learning, which necessitates labelled datasets that are difficult to obtain. Crowd-sourcing is the most commonly used method for constructing datasets, which is not ideal for evaluating machine learning techniques in real-world scenarios. Therefore, it is preferable to evaluate classifiers on real-world applications to develop algorithms that work efficiently in practical settings. This comprehensive review is valuable to researchers seeking an in-depth clarity on this field's key components. This highlights significant advancements and suggests future directions.

REFERENCES

[1] E. P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, H. W. Lauw, "Detecting product review spammers using rating behaviors", Oct.2010.

[2] Ott, Myle & Choi, Yejin & Cardie, Claire & Hancock, Jeffrey. "Finding Deceptive Opinion Spam by Any Stretch of the Imagination", 2011.

[3] J. K. Rout, A. Dalmia, and K.-K.R. Choo, Revisiting semi -supervised learning for online deceptive review detection, ‖ IEEE Access, Vol. 5, pp.1319–1327, 2017.

[4] E. F. Cardoso, R. M. Silva, and T. A. Almeida, ``Towards automatic altering of fake reviews," Neurocomputing, vol. 309, pp. 106116, Oct. 2018.

[5] L. Da Xu, W. He, and S. Li, ``Internet of Things in industries: A survey," IEEE Trans. Ind. Informat., vol. 10, no. 4, pp.

22332243, Nov. 2014.
[6] Y. Ren and Y. Zhang, “Deceptive opinion spam detection using neural network”, in Proc. 26th Int. Conf. Compute. Linguistics: Tech. Papers (COLING), 2016, pp. 140150.
[7] N. Jindal and B. Liu, “Opinion spam and analysis”, in Proc. Int. Conf. Web Search Web Data Mining (WSDM), 2008, pp. 219230.
[8] A. Heydari, M. Tavakoli, and N. Salim, “Detection of fake opinions using time series”, Expert Syst. Appl., vol. 58, pp. 8392, Oct. 2016.
[9] L. Li, W. Ren, B. Qin, and T. Liu, “Learning document representation for deceptive opinion spam detection,” in Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Nanjing, China: Springer, 2015, pp. 393404.
[10] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna, “Detecting deceptive reviews using generative adversarial networks”, in Proc. IEEE Secure. Privacy Workshops (SPW), May 2018, pp. 8995.
[11] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “Fake review detection: Classification and analysis of real and pseudo reviews”, Univ. Illinois Chicago, Chicago, IL, USA, Tech. Rep. UIC-CS-03-2013, 2013.
[12] R. Yafeng, J. Donghong, Z. Hongbin, and Y. Lan, “Deceptive reviews detection based on positive and unlabeled learning”, J. Compute. Res. Develop., vol. 52, no. 3, p. 639, 2015.
[13] R. Y. K. Lau, S. Y. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, “Text mining and probabilistic language modeling for online review spam detection”, ACM Trans. Manage. Inf. Syst., vol. 2, no. 4, pp. 130, Dec. 2011.
[14] P. K. Novak, J. Smailović, B. Sluban and I. Mozeti, “Sentiment of Emojis,” Journal of
Computation and Language, vol.10, no. 12, pp. 1-4, December 2015.