

# An Approach to Detect Fake Reviews based on Logistic Regression using Review-Centric Features

Kaushik Daiv<sup>1</sup>, Mrunal Lachake<sup>2</sup>, Prathamesh Jagtap<sup>3</sup>, Srishti Dhariwal<sup>4</sup>, Prof. Vitthal Gutte<sup>5</sup>

<sup>1</sup>Kaushik Daiv, Computer Science, MITCOE, Pune, India

<sup>2</sup>Mrunal Lachake, Computer Science, MITCOE, Pune, India

<sup>3</sup>Prathamesh Jagtap, Information Technology, MITCOE, Pune, India

<sup>4</sup>Srishti Dhariwal, Computer Science, MITCOE, Pune, India

<sup>5</sup>Prof. Vitthal Gutte, Information Technology, MITCOE, Pune, India

\*\*\*

**Abstract** - The impact of reviews on any e-commerce site is of great importance, as it can be the base for a buyer's decision to buy any product. Buyer tries to evaluate the authenticity and quality of the product using the feedback given by other previous buyers in the form of review. But, sellers are taking advantage of the reviews by posting reviews in an attempt to promote or defame a product. Such reviews which are not a genuine opinion of an individual are termed as fake reviews. The existence of such fake reviews makes the buyer unable to make the right judgments of sellers, which can also cause the credibility of the platform to be downgraded. Thus, it is very important to identify the fake reviews on the platform. In this paper, we propose a method to detect such fake reviews using a logistic regression model by considering review centric features achieving an overall accuracy of 82%. Also, our study illustrates the impact of the "verified purchase" feature in fake review classification.

**Key Words:** CountVectorizer, Fake review, Feature extraction, Logistic Regression, NLP, Tf-idf.

## 1. INTRODUCTION

In this era of the Internet, people can easily share their views about products and services by using e-commerce sites, forums, and blogs. Consumers tend to refer to reviews of other consumers to buy products on e-commerce websites. These reviews are helpful for potential customers and vendors too. Vendors are also capable of designing their additional marketing strategies based on the consumers' reviews. For example, if various consumers buy a specific model of a laptop and write reviews regarding issues concerning its screen resolution or processor speed, then vendors might contact associated manufacturers and make them aware of their issues and resolve them in order to increase customer satisfaction towards their products or services.

Targeted products or services may be promoted or degraded by mischievous users by writing either applaudable reviews or derogatory reviews. Therefore the integrity of such reviews is questionable [2]. Such reviews are known as fake reviews. Recently the media news from the New York Times and BBC has stated that "nowadays,

fake reviews are very frequent on websites, and recently a photography company was exposed to thousands of customer fake reviews". Also it has been reported that 88% of consumers trust online reviews as much as personal recommendations [6]. Hence, detecting fake reviews appears to be a key area, and without solving this important issue, online review sites could become a place full of lies, almost rendering the e-commerce business useless. To counter this issue, some researchers have already made some progress in detecting fake reviews that we discuss in the next section. However, most of the previous research was done on hotel reviews and there are a lot of labeled datasets available for hotel reviews including Yelp. But there is still room for improvement in the detection of product reviews and hence, in this paper we are focussing on product reviews provided by amazon using supervised learning techniques as supervised models give more accurate results than unsupervised or semi-supervised models [8].

This paper proposes a method for detecting fake reviews for Amazon products using a statistics model called logistic regression by employing Term frequency - inverse document frequency (Tf-idf) and CountVectorizer for feature extraction. In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. (In our case, whether it is a fake review or a genuine review). Each review would be assigned a probability between 0 and 1. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. It is a form of binary regression. In a binary logistic regression model, the dependent variable has two labels (categorical). The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though in our paper, we are using it as a classifier for fake review classification, by choosing a cutoff value of 0.5 and classifying inputs with probability greater than the cutoff as fake reviews, while below the cutoff as the genuine reviews; this is a common way to make a binary classifier. In addition to this, our study proposes the use of a "verified purchase" feature (refer to Section 3.4 for

details) with a logistic regression model for fake review classification.

There is a lot of research employing logistic regression in the field of fake review detection which considers various features like review features, reviewer features, topic features, etc. Our study is unique as we extend their research by taking into account “verified purchase” feature which we discuss in section 3.4. This study might be helpful for future researchers who might want to improve the fake review detection system.

The two main contributions of this study are as follows: (i) Provide researchers and practitioners with insight and further improvement prospects on the fake review detection problem based on logistic regression model using review centric features. (ii) Present the effect of using the “verified purchase” feature for identifying the fake reviews .

The remaining paper is divided into the following sections: Section 2 discusses the related work to fake review detection, logistic regression, and feature extraction. Section 3 states the problem definition. Section 4 gives an insight of the dataset used. Section 5 describes the detailed methodology and feature extraction techniques used in our study. Section 6 discusses the experiments and results. Finally, in Section 7 we present our final conclusions and Section 8 suggests possible areas for future work.

## 2. RELATED WORK

In recent years, researchers have studied a lot about fake reviews and their detection techniques involving machine learning paradigms. The research shows that detecting fake reviews is not only limited to the review content, but it is evident that reviewers’ behaviour also plays a big role in detecting the fakeness of the review.

While there are various machine learning techniques viz supervised learning, semi-supervised learning, unsupervised learning techniques; in [8], researchers used supervised machine learning algorithms (Logistic Regression, K-Nearest Neighbor, Naive Bayes) on Yelp dataset.

A study by Faliang Huang et al.[4] proposed Ensemble learning is a powerful machine learning paradigm that has exhibited apparent advantages in many applications. An ensemble in the context of machine learning can be broadly defined as a machine learning system that is constructed with a set of individual models working in parallel and whose outputs are combined with a decision fusion strategy to produce a single answer for a given problem. Their study states that the idea of ensemble learning is principally based on the theory foundation stone that the generalization ability of an ensemble is usually much stronger than that of a single learner. So this application proves helpful to boost the weak learner into a strong one.

While the ensemble learning did show a new way to approach this detection problem, Brian Heredia et al. [3] suggested a way to improve ensemble learning by employing ensemble learning with feature selection techniques. They employed three of these techniques: Select-Boost, Select-Bagging, and Random Forest. The study shows that applying a combination of ensemble and feature selection (Select-Boost) has significant improvement in performance when compared to using solely MNB. The results indicate the combination of Select-Boost, MNB, and Chi-Squared (or signal-to-noise) to be the best performing model, significantly outperforming all other methods, with the exception of RF500.

The way to approach through an algorithm can be different, which was given a base by Fathima Nadal et al. [16] as they stated that the existing system and research work reveal that most classification algorithms perform well to detect or predict the fakeness of any article. Though logistic regression serves well for this purpose, our system is based on this information and thus we focus to work with classification algorithms like logistic regression.

## 3. PROBLEM DEFINITION

To design and build an efficient and robust fake review detecting model for e-commerce product reviews using a logistic regression model which follows the idea of modelling the probability for classifying fake and genuine reviews. Also provide a consensus strategy for feature extraction and text preprocessing.

## 4. DATASET

The dataset used in this paper is “amazon\_reviews” which is openly available on Kaggle. The dataset contains Amazon product reviews which are labelled as `_label1_` and `_label2_` as fake review and genuine review respectively. The dataset contains a total of 21,000 uniformly distributed reviews in which 50% are fake reviews and 50% are genuine reviews. The dataset contains different columns viz. rating, verified purchase, product category, product id, product title along with review content and its label. You can find the full dataset by the link provided in ref. [15].

## 5. PROPOSED METHOD

This section elaborates the proposed methodology for the above-stated problem definition. We divide our methodology in six steps as follows:

1. Collect the data of labelled reviews.
2. Preprocessing the dataset.
3. Feature extraction.
4. Training the logistic regression model and hyperparameter tuning.
5. Evaluating the model on Test data.

First step is to collect the labelled dataset of reviews. There are multiple datasets available online which are used in previous research, however finding a labelled dataset for reviews was a difficult task. Fortunately, we were able to find a labelled dataset on kaggle provided by Amazon. The dataset contains a total of 21,000 reviews in which 50% are fake reviews and 50% are genuine reviews.

After finding the dataset, the next step involves preprocessing the data. The dataset cannot be directly used to train the classifier model as the model cannot handle the text data. Preprocessing includes removing stop words and punctuations, stemming, lemmatization, etc. Preprocessing is discussed in the next section in detail.

In the feature extraction phase, Tf-idf(Term Frequency and Inverse Document Frequency) and CountVectorizer were used. Tf-idf increases the weight of uncommon words and decreases the weight of common words and results in creating a vector of features [1]. CountVectorizer creates a vector of features based on the frequency of the words in the review. Our research shows that CountVectorizer outperforms Tf-idf which may be because the dataset has reviews on many different products resulting in many uncommon words.

Followed by the feature extraction phase, we trained our logistic regression model on the set of features selected in the feature extraction phase. Training the model is just not sufficient to classify the reviews successfully. For better accuracy we need to tune the hyperparameters of the model viz. C, penalty, solver, etc. The process of hyperparameter tuning is discussed in detail in the Experiment section.

After the model is trained and tuned, the model needs to be evaluated to understand its performance. Evaluation of the model is done by testing it on unlabelled data and

calculating the accuracy, precision and recall. Results are discussed in the Experiment section.

### 5.1 DATA PREPROCESSING AND FEATURE EXTRACTION

Variety of features that have been proposed and used separately by supervised approaches to identify fake reviews; which are review centric and reviewer centric features. In some cases review-centric features are considered separately. In other cases reviewer-centric features are taken into account. From the previous research proposed in [8], [14], we tried to pick the best features which would help identify the fake reviews. We have considered the following:

#### 1. Rating

Users can rate the product from 1 to 5 stars representing satisfaction/dissatisfaction about the product. This feature can be used to validate that the review written and the ratings given by the reviewer are intended only in one direction and do not contradict. Also, ratings corresponding to the fake reviews usually deviate from the average rating of the product[8]. Thus, helping in classifying the fake reviews.

#### 2. Verified purchase

The verified purchase means Amazon has verified that the person writing the review has purchased the product at Amazon and didn't receive the product at deep discount. This feature helps to consider those reviews which are genuine as we get to know which purchaser has actually bought the product and used it.

#### 3. Review length

The length of the review is also considered for training the model as the previous research by Xinyue Wang et al.[8] suggests that reviews written by spammers are very short and defame/promote the product.

Along with the above two review centric features, review content is also considered for classification. The review text needs to be processed before passing to the model. The first step of processing includes removal of all the characters and expressions other than letters as the model can't make any sense of the punctuations and expressions. The review text is then split into a list of words and then each word is converted to its base form by stemming; followed by removal of stopwords. Stopwords are the frequently occurring words useful syntactically and grammatically that do not add any value to the model. A corpus of these words is generated. The CountVectorizer function provided by sklearn in python is used to represent the corpus of words using a sparse matrix where each word acts as a column and the review as a row having the most frequent 1400 words from the corpus. This sparse matrix of 1400 most frequent words is used as a feature vector to the model along with the verified

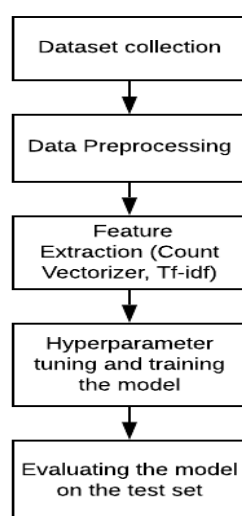


Figure -1: Block diagram for proposed logistic regression model

purchase, rating and review length of the product. Similarly, we also create a feature vector of 1400 words using Tf-idf(Term Frequency and Inverse Document Frequency).

### 6. EXPERIMENTS AND RESULTS

This section describes the experiments in our research. Dataset cannot be passed directly for training. At first relevant feature columns were extracted like rating, verified purchase. Then, categorical features were converted to numerical features using LabelEncoder. Text data was extracted using CountVectorizer and Tf-idf (max\_features = 1400). This was followed by calculating length for each review. Finally, our data having a feature vector of 1403 (1400 features of review content, rating, verified purchase, review length) was generated. Train test split of 80-20% was carried out on the data. This was followed by scaling the data by using MinMaxScaler class from sklearn library. The MinMaxScaler transforms features by scaling each feature to a given range. This range can be set by specifying the feature\_range parameter (default at (0,1)).

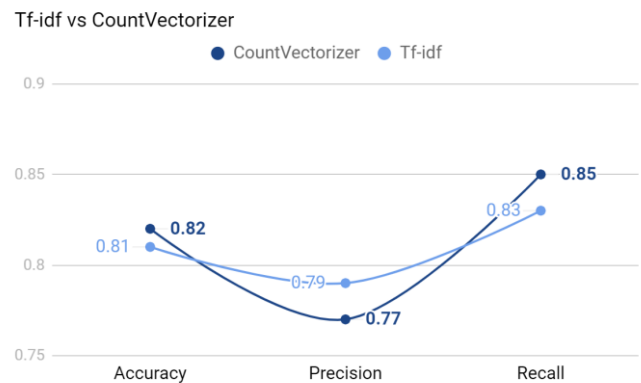
Logistic regression model was trained on the train data for both CountVectorizer and Tf-idf using LogisticRegression in sklearn. Followed by this, hyperparameters of the model were tuned using GridSearch. Following is the list of optimal hyperparameters that result in best accuracy: c = 1, penalty = 'l2', solver = 'newton-cg'.

Also, to examine the impact of using "verified purchase" as a feature for fake review classification, the XGBoost with CountVectorizer with the same hyperparameters was trained without using "verified purchase" as a feature and the model was evaluated for the performance.

The last phase of the experiment is evaluation of the model to understand the performance. Binary classification involves classifying the data in two groups eg. yes/no, true/false, fake/genuine, etc. Target variables in such problems are not continuous but they predict the probabilities to be yes/no, etc. Such models are evaluated using a metric called confusion matrix. Using confusion matrix, we have calculated accuracy, precision and recall for both logistic regression(LR) with CountVectorizer and logistic regression with Tf-idf; with and without "verified purchase" feature. It was evident from the results that the accuracy of the CountVectorizer exceeded by a smaller margin than Tf-idf as shown below in Table 1.

| Classifiers             | Accuracy     | Precision    | Recall       |
|-------------------------|--------------|--------------|--------------|
| LR with Tf-idf          | 0.818        | 0.794        | 0.834        |
| LR with CountVectorizer | <b>0.821</b> | <b>0.773</b> | <b>0.855</b> |

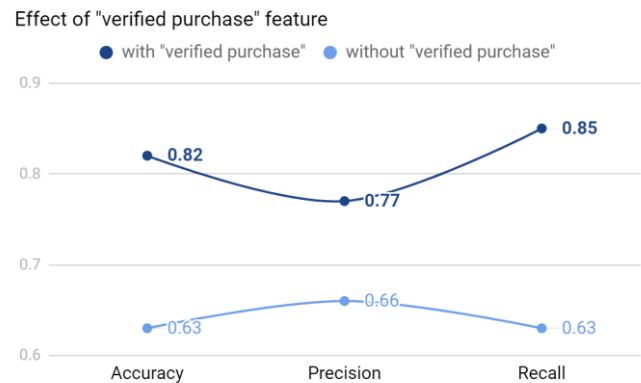
**Table -1:** Performance of logistic regression with Tf-idf and CountVectorizer



**Figure -2:** Accuracy, precision, recall comparison - Tf-idf vs CountVectorizer using logistic regression

| Classifiers                    | Accuracy     | Precision    | Recall       |
|--------------------------------|--------------|--------------|--------------|
| LR without "verified purchase" | 0.639        | 0.669        | 0.631        |
| LR with "verified purchase"    | <b>0.821</b> | <b>0.773</b> | <b>0.855</b> |

**Table -2:** Performance of logistic regression with and without "verified purchase" feature



**Figure -3:** Accuracy, precision and recall comparison - with "verified purchase" vs without "verified purchase"

Our results from Table 1 and Fig. 2 show that the logistic regression model with Tf-idf and CountVectorizer using review-centric features performs extremely well on our dataset by achieving an accuracy of 81% and 82% respectively. 1% of test data is around 400 reviews. Thus, it is evident that CountVectorizer classifies around 400 reviews more correctly than Tf-idf. Though, accuracy of CountVectorizer feature extraction technique outperforms Tf-idf by a margin of 1%; precision of Tf-idf feature exceeds CountVectorizer by 2%. Table 1 and Fig. 2 show the accuracy, precision and recall of Tf-idf and CountVectorizer.

We also observe that, “verified purchase” feature has a significant impact on fake reviews classification. Fig. 3 and Table 2 show the accuracy of logistic regression with CountVectorizer using “verified purchase” feature is 82% whereas when the feature is not used for classification, the same model achieves an accuracy of 63%. This result suggests that the “verified purchase” feature is very effective for identifying fake reviews. Table 2 and Fig. 3 show the accuracy, precision and recall of logistic regression using CountVectorizer with “verified purchase” and without “verified purchase”.

## 7. CONCLUSIONS

In this paper, we have discussed the impact of logistic regression model for identifying the fake reviews using review-centric features. Along with review content, we have provided a set of review-centric features for classification of the fake reviews. One of the review-centric features we propose in this paper is “verified purchase”. Our research shows that using “verified purchase” as a feature for classifying fake reviews has an outstanding effect. In addition to this, we have proposed two feature extraction techniques viz. Tf-idf and CountVectorizer and therefore, conclude that implementing logistic regression with CountVectorizer on the used dataset has achieved an accuracy of 82% and an accuracy of 81% with Tf-idf. The work proposed in this paper acts as a platform for further research in logistic regression in fake review detection. This study might be helpful for future researchers who want to improve the fake review detection system using logistic regression. The use of “verified purchase” feature for the classification is a prominent contribution of our research to the domain of fake review classification.

## 8. FUTURE SCOPE

This paper has presented the effect of feature “verified purchase” on the result, so further research can consider this feature for classification. This paper focuses only on review-centric features, future researchers can work on reviewer-centric features with logistic regression. Additionally, future work may involve testing this process on other data sets to see if results generalize. In future, we try to increase the accuracy of logistic regression by exploring other effective features and parameters and try to build a better model. Also, future studies may try to generalize a model for multilingual reviews and might try considering the sentiment of the emoticons used in the reviews.

## REFERENCES

- [1] Sifat Ahmed, Faisal Muhammad Shah, “Using Boosting Approaches to Detect Spam Reviews”, 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology(ICASERT), 2019.
- [2] Faisal Khurshid, Yan Zhu, Zhuang Xu, Mushtaq Ahmad, Muqet Ahmad, “Enactment of Ensemble learning for

Review Spam Detection on Selected Features”. International Journal of Computational Intelligence System 2019, Vol.12(1); pp.387-394; ISSN: 1875-6891; 2019.for every journal there is volume and issue number, please check for all papers from this paper onwards and make necessary changes, thanks

- [3] Brian Heredia, Taghi M. Khoshgoftaar, Joseph D. Prusa, Michael Crawford, “Improving detection of untrustworthy online reviews using ensemble learners combined with feature selection”, Springer, Article number: 37, 2017.
- [4] Faliang Huang, Guoqing Xie, Ruliang Xiao, “Research on Ensemble Learning”, International Conference on Artificial Intelligence and Computational Intelligence, 2009.
- [5] Dushyanthi U. Vidanagama, Thushari P. Silva, Asoka S. Karunananda, “Deceptive consumer review detection: a survey”, Springer 2019. Available at <https://doi.org/10.1007/s10462-019-09697-5>
- [6] Ioannis Dematis, Eirini Karapistoli, Athena Vakali, “Fake Review Detection via Exploitation of Spam Indicators and Reviewer Behavior Characteristics”. Springer International Publishing AG pp.581-595, 2018.
- [7] Julien Fontanarava, Gabriella Pasi, Macro viviani, “Feature Analysis for Fake Review Detection through Supervised Classification”. IEEE 2017 International Conference on Data science and analytic, 2017.
- [8] Xinyue Wang, Xianguo Zhang, Chengzhi Jiang, Haihang Liu, “Identification of Fake Reviews Using Semantic and Behavioral Features”, 2018 4th IEEE International Conference on Information Management, 2018.
- [9] Anna V. Sandifer, Casey Wilson, Aspen Olmsted, “Detection of fake online hotel reviews” ,IEEE Internet Technology and Secured Transactions (ICITST), International Conference, 2017.
- [10] Amitkumar B. Jadhav, Vijay U. Rathod and Dr. Hemantkumar B. Jadhav, “Improving Performance of Fake Reviews Detection in Online Review's using Semi-Supervised Learning”. International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 06 | June 2019.
- [11] Min-Yuh, Chih-Chien Wang, Chien-Chang Chen and Shao-Chieh Yang, “Exploring Review Spammers by Review Similarity: A Case of Fake Review in Taiwan”. Proceedings of the Third International Conference on Electronics and Software Science (ICESS2017), Takamatsu, Japan, 2017.
- [12] Xian Wu, Yuxiao Dong, Jun Tao, Chao Huang and Nitesh V. Chawla, “Reliable Fake Review Detection via Modeling Temporal and Behavioral Patterns”. 2017 IEEE International Conference on Big Data (BIG DATA), 2017.
- [13] SP. Rajamohana, Dr. K. Umamaheswari, M. Dharani and R. Vedackshya, “A Survey On Online Review Spam Detection Techniques’ . IEEE International Conference

on Innovations in Green Energy and Healthcare Technologies (ICIGEHT'17), 2017.

- [14] Huaxun Deng, Linfeng Zhao, Ning Luo, Yuan Liu, Guibing Guo, Xingwei Wang, Zhenhua Tan, Shuang Wang and Fucai Zhou, "Semi-supervised Learning based Fake Review Detection", 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), 2017.
- [15] Amazon labelled dataset for fake product reviews used in this paper is available on: <https://www.kaggle.com/lievgarciya/amazon-reviews>.
- [16] Fathima Nada, Bariya Firdous Khan, Aroofa Maryam, Nooruz-Zuha, Zameer Ahmed. "Fake news detection using Logistic Regression." International Research Journal of Engineering and Technology (2019): p-ISSN: 2395-0072
- [17] Muhammad Syahmi Mokhtar, Yusmadi Yah Jusoh, Novia Admodisastro, Noraini Che Pa, Amru Yusrin Amruddin. "Fakebuster: Fake News Detection System Using Logistic Regression Technique in Machine Learning." International Journal of Engineering and Advanced Technology (2019): ISSN: 2249 – 8958.
- [18] Pankaj Chaudhary, Abhimanyu Tyagi, Santosh Mishra. "Fake Review Detection through Supervised Classification." Indian Conference on Recent Innovations in Emerging Technology & Science (2018):ISSN: 2320-2882