International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

An AI-Driven Method for Detecting Fake Reviews through Feature Engineering

Dr Vuppu Padmakar[1], Dr B V Ramana Murthy[2], Dr DVSS Subrahmanyam[3]

[1]Associate Professor, Dept. of CSE (AIML) Neil Gogte Institute of Technology, Hyderabad
[2]Professor & Vice Principal, Dept. of CSE, Stanley College of Engineering & Technology for Women, Hyderabad
[3]Professor & Vice Principal, Keshav Memorial Engineering College, Hyderabad

Abstract:
This research aims to develop a model capable of distinguishing between genuine and fraudulent reviews, thereby assisting customers in avoiding online scams. Businesses also stand to gain, as enhanced trust can lead to increased sales. The study focuses on refining the prediction system for identifying fake reviews by utilizing real-time datasets from Amazon to train the model. Various machine learning algorithms, including Random Forest, AdaBoost, and Naïve Bayes, will be employed for classification purposes. The effectiveness of each algorithm will be evaluated using a confusion matrix. A detection process will be implemented to ascertain the authenticity of reviews through feature engineering. By leveraging Natural Language Processing (NLP) to extract significant features from the text, the research will facilitate the detection of review spam.

Keywords: Review, Feedback, AdaBoost, Naïve bayes, Random Forest

1. Introduction:

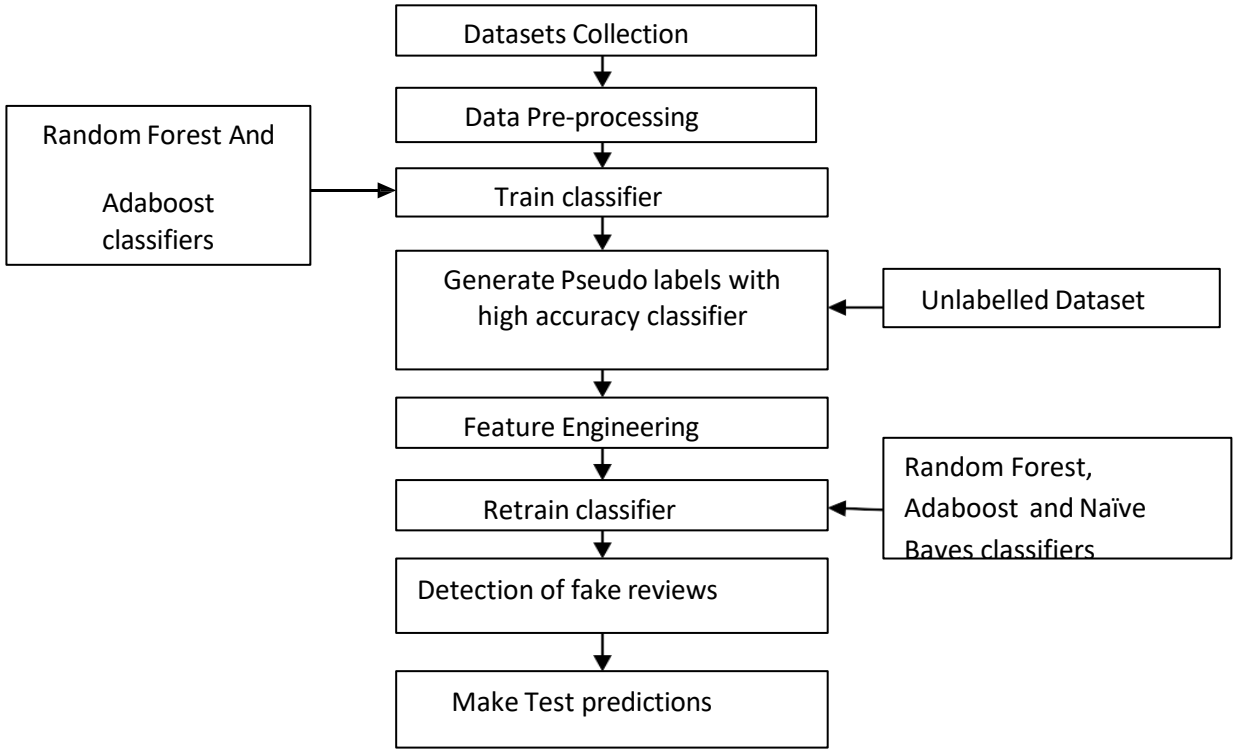
Online reviews provided by customers significantly influence their purchasing decisions and serve as a crucial source of information for assessing public sentiment regarding products or services. Customers often seek to gauge the authenticity and quality of a product by examining feedback from previous buyers in the form of reviews. Given the substantial impact of these reviews on consumer behaviour, manufacturers and retailers are increasingly attentive to customer feedback and reviews. This reliance on online reviews has raised concerns about the potential for fraudulent activities, where individuals may generate fake reviews to artificially enhance or diminish the reputation of products and services. This phenomenon is referred to as Review Spam, wherein spammers exploit reviews for financial gain. By extracting meaningful features from these reviews, it is feasible to implement review spam detection through various machine learning methodologies. Data mining and machine learning techniques play a pivotal role in identifying fraudulent reviews. This paper utilizes two datasets: one labeled dataset from Amazon and another unlabeled dataset. Unlike the AMT dataset, which contains duplicate reviews, these datasets feature authentic reviews. Data preprocessing has been conducted on both datasets. The unlabeled dataset is extensive, making manual labeling impractical. It also includes additional attributes

such as product ratings and review images, which assist in distinguishing between genuine and fake reviews. Feature engineering is employed to extract relevant features from the data.

The attributes such as Verified Purchase included in the dataset significantly influence the ability to discern whether a review is authentic or fraudulent. The product rating within the dataset aids in assessing the deviation in ratings. Additionally, the presence of a review image indicates whether the customer has uploaded a picture of the product. If a review from a non-verified purchaser includes an image of the product, it may be regarded as a legitimate review. These behavioural characteristics of the reviews contribute to a more accurate classification of genuine and deceptive reviews. To label the unannotated data, a self-training approach, which is a semi-supervised learning technique, is employed. Initially, supervised learning is conducted on the labeled dataset utilizing algorithms such as AdaBoost, Naïve Bayes, and Random Forest, with the Random Forest algorithm yielding the highest accuracy. Subsequently, self-training is performed using Random Forest to label the unannotated Amazon dataset sourced directly from the website. The features are then applied to the newly labeled dataset, allowing for the classification of reviews as either real or fake.

2. Proposed System

The proposed system aims to differentiate between genuine and fraudulent reviews. To achieve this classification, a machine learning (ML) model requires training, which necessitates the use of datasets. The system incorporates two distinct datasets for training the ML model and employs three different classification algorithms for the classification process. Both datasets are sourced from real-life Amazon data. The first dataset is a labeled set with limited features for review classification, while the second dataset, which contains a broader range of features beneficial for classification, is unlabeled. To label this second dataset, a semi-supervised algorithm is employed. In this paper, self-training is utilized to annotate the second dataset, which includes reviews accompanied by images as well as those without. The images are regarded as a significant feature in determining the authenticity of a review. Additionally, rating deviation and verified purchase status are also critical features that contribute to the classification of the reviews.



3. Datasets Collection

Acquiring the appropriate data in the correct format presents one of the most significant challenges in the field of machine learning. This process involves gathering or identifying data that is correlated with the desired outcomes, specifically data that provides insights into the events of interest. It is essential that the data is pertinent to the issue at hand; for instance, in the development of a facial recognition system, images of kittens would not be useful. A data scientist must ensure that the data aligns with the specific problem being addressed. In the absence of the requisite data, one must revert to the data collection phase before proceeding with the design of an AI solution.

Machine learning relies on a well-structured training set to yield accurate results. The process of assembling and creating this training set, which consists of a substantial amount of known data, requires considerable time and specialized knowledge regarding where to find valuable data. The training set acts as a reference point for deep learning networks during their training phase, enabling them to learn before they are tasked with analyzing unfamiliar data. At this stage, knowledgeable individuals must identify the appropriate raw data and transform it into a tensor, a numerical format that deep learning algorithms can interpret. In many respects, the assembly of a training set resembles a preparatory phase prior to the actual training. In this module, we have compiled two distinct datasets containing reviews for various products. The datasets are as follows,

Dataset 1: Amazon Labeled Data Set
Dataset 2: Amazon Unlabeled Data Set

4. Feature Engineering:

Feature selection techniques are employed to identify and remove undesirable, irrelevant, and redundant attributes from datasets that do not enhance the performance of a predictive model or may even detract from its accuracy. This concept is essential in machine learning, significantly influencing the efficiency of your model. The choice of data attributes for training machine learning models plays a crucial role in determining the outcomes. The presence of irrelevant or only partially relevant features can adversely affect model performance. The primary objective of feature selection is to enhance accuracy while minimizing training time. This process involves selecting features that have a greater contribution to the target variable or output of interest, either through automated methods or manual selection. Irrelevant characteristics in the dataset can diminish model accuracy and lead to training based on non-essential features.

Advantages of conducting feature selection prior to modeling your data include:

Mitigating Overfitting: A reduction in redundant data decreases the likelihood of making decisions influenced by noise.

Enhancing Accuracy: The accuracy of the model improves due to the elimination of misleading data.

Decreasing Training Time: A smaller number of data points simplifies algorithm complexity, resulting in faster training times. In this paper, feature selection is performed manually; for instance, attributes such as reviewer name and review date are excluded as they negatively impacted accuracy. Feature extraction, a critical phase in the pattern recognition or machine learning workflow, aims to enhance performance. This process involves distilling data to its most significant elements, thereby providing more valuable information to machine and deep learning models. It primarily focuses on removing extraneous features that could influence predictive performance. Various methodologies have been proposed in the literature for extracting features relevant to fake review detection, with one prevalent approach being the utilization of textual features. The textual feature method known as TF-IDF calculates the frequency of both true and false documents (TF) alongside the inverse document frequency (IDF). Each term is assigned its own TF and IDF score, and the TF-IDF weight for a term is determined by multiplying its TF and IDF values.

The reviews are categorized into four distinct groups utilizing a confusion matrix:
True Negative (TN): where genuine events are accurately identified as genuine,
True Positive (TP): where fraudulent events are correctly identified as fraudulent,
False Positive (FP): where genuine events are incorrectly labeled as fraudulent, and
False Negative (FN): where fraudulent events are mistakenly classified as genuine.

Additionally, the analysis incorporates the user's personal profile and behavioral characteristics. Two primary methods for identifying spammers are employed: one involves examining the frequency and uniqueness of a user's comment timestamps compared to typical users, and the other assesses whether a user submits repetitive similar reviews that lack relevance to the target domain. In this study, we utilize

TF-IDF to extract features from the content in a unigrams format.

Feature Engineering: Fake reviews exhibit various descriptive characteristics associated with the behaviors of reviewers while composing their evaluations. In this paper, we examine several of these characteristics and their influence on the effectiveness of the fake review detection process. Specifically, we focus on rating deviation, verified purchases, and review images as behavioral indicators.

Rating: Users have the ability to evaluate the product on a scale of 1 to 5 stars, reflecting their level of satisfaction or dissatisfaction. This functionality serves to confirm that the reviews and ratings provided by the reviewer are consistent and do not contradict one another. Additionally, ratings associated with fraudulent reviews typically diverge from the product's average rating, thereby assisting in the identification of such reviews.

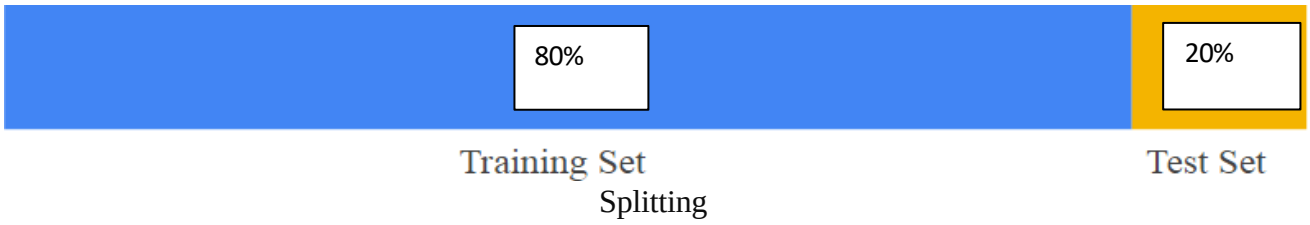
Verified purchase: A verified purchase indicates that Amazon has confirmed the reviewer has indeed bought the product from their platform and did not acquire it at a significant discount. This feature is instrumental in distinguishing genuine reviews, as it clarifies which customers have legitimately purchased and utilized the product.

Review image: The inclusion of a review image is a crucial aspect that aids in verifying whether the reviewer has genuinely acquired the product and subsequently provided a review. When a reviewer includes an image of the product alongside their review, it is regarded as a credible assessment.

There exist both verified purchase reviews and non-verified purchase reviews. Verified purchase reviews are marked with a "verified purchase" label, signifying that the customer has bought the product without any substantial discount. However, this label is only assigned if the customer has a purchase history totaling $50 (approximately 4000 INR) within the last twelve months. Consequently, non-verified purchase reviews may also be authentic. To ascertain the legitimacy of non-verified purchase reviews, we can rely on images submitted by customers. If a customer includes an image, it is deemed a legitimate review.

5. Data Splitting

The model could not be adequately trained on the available data in the context of machine learning, which prevents us from asserting its accuracy on real-world data. To address this issue, it is essential to ensure that our model has successfully identified the appropriate patterns within the dataset. We conduct data splitting to assess whether the class values predicted by the machine learning model align with the actual class values. The evaluation results provide insight into the performance rate of our machine learning model. In this study, I am utilizing three distinct algorithms: Random Forest, AdaBoost, and Naïve Bayes.



Splitting

Splitting 80:20: I have noted that the datasets contain instances from the same class, which may lead to bias in the algorithm and a subsequent decline in its performance. To address this issue, I have randomly sampled data from the dataset and divided it into an 80:20 ratio. Specifically, 80% of the data is allocated for training the model, while the remaining 20% is designated for testing. This approach is widely recognized for its simplicity and effectiveness in yielding a less biased or overly optimistic estimate of the model's performance.

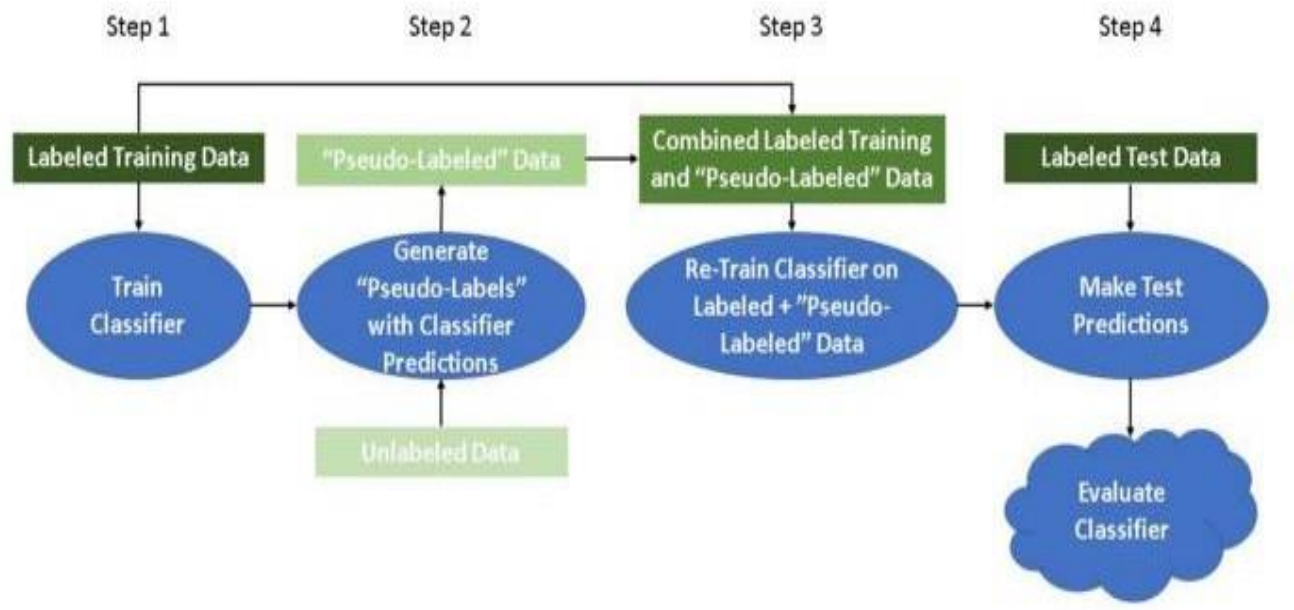
Benefits of a train/test split include:
K-fold cross-validation operates K times more efficiently than traditional cross-validation due to the repetition of the train/test split K times. This approach facilitates a more thorough examination of the testing process's results. The representation of data in a visual format is referred to as data visualization. This technique enhances data understanding by condensing and displaying extensive information in a straightforward and accessible manner, thereby promoting clear and effective communication of insights.

Building ML Model: The process of training a machine learning algorithm involves supplying training data to the learning algorithm. The outcome of this training is known as an ML model. This model is capable of making predictions on new data where the target variable is not known. It is essential that the training data includes the correct label, often referred to as the target or target attribute. The learning algorithm analyzes the training data to identify patterns that link the input attributes to the desired output, ultimately producing an ML model that encapsulates these relationships. This model can then be utilized to predict outcomes for new data with unknown targets. In this paper, we will develop Random Forest, AdaBoost, and Naïve Bayes models for the classification of the data. Our dataset requires binary classification ML models, as the classes within the datasets are represented in binary format, such as real or fake.

Semi-supervised: In the realm of machine learning classification tasks, the availability of extensive data for training algorithms significantly enhances performance. In supervised learning, it is essential that the data utilized is appropriately labeled in relation to the target class; otherwise, the algorithms will struggle to discern the connections between the target and independent variables. When developing large, labeled datasets for classification purposes, several considerations must be taken into account:

1. The process of data labeling can be time-consuming.
2. The expenses associated with data labeling can be substantial.

If resources such as time and budget are limited, resulting in only a portion of a large dataset being labeled, the remaining unlabeled data can be addressed through semi-supervised learning. This approach involves training a classifier on a small amount of labeled data, which can subsequently be employed to predict outcomes for the unlabeled data. The predictions made for the unlabeled data can serve as "pseudo-labels" in subsequent iterations of the classifier, as they are likely to be more accurate than random guesses. Among the various methodologies of semi-supervised learning, self-training is a notable example.



Self-Training

Self-training at a conceptual level is conducted as follows:

Step 1: Initially, partition the labeled data instances into training and testing sets. Subsequently, utilize the labeled training data to train a classification algorithm.

Step 2: Utilize the trained classifier to assign class labels to all unlabeled data. The 'pseudo-labels' are selected from the predicted class labels that exhibit the highest probability of accuracy.

(There are several alternatives for Step 2: a) All predicted labels may be accepted as 'pseudo-labels' simultaneously, irrespective of their probability, or b) The 'pseudo-labeled' data can be adjusted based on the confidence of the predictions.)

Step 3: Merge the 'pseudo-labeled' data with the labeled training data. With this combined dataset, retrain the classifier.

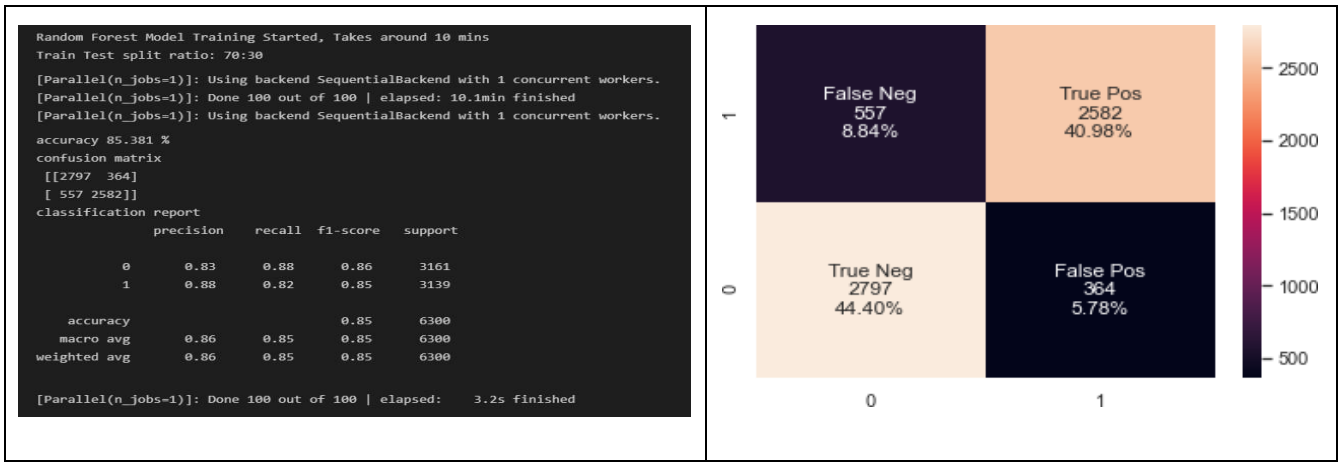
Step 4: Employ the trained classifier to predict class labels for the labeled test data instances. Utilize your chosen metrics to evaluate the performance of the classifier.

(Repeat Steps 1–4 until no further predicted class labels from Step 2 meet a specified probability threshold, or until all unlabeled data has been processed.)

Results Analysis:

Random forest Algorithm

Dataset 1: Amazon Labelled Reviews Dataset

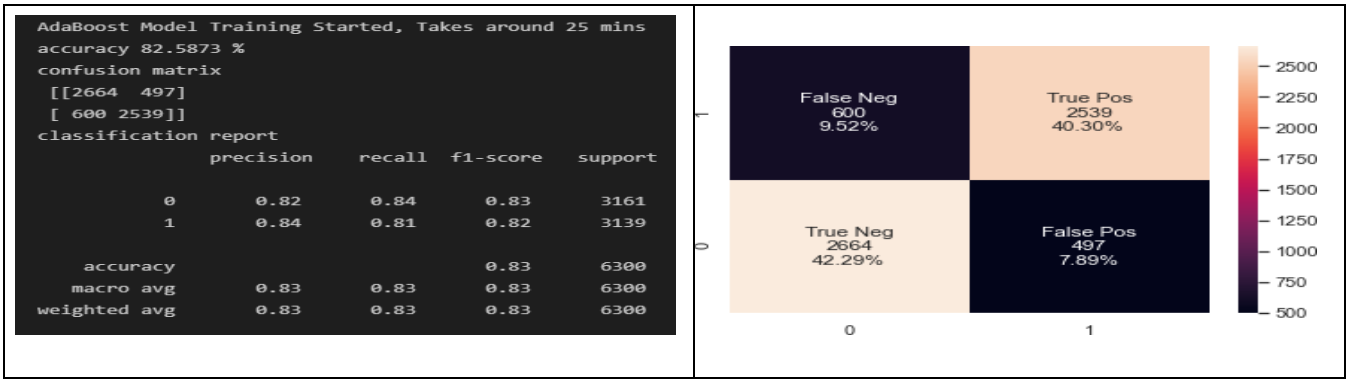


Random forest accuracy dataset 1

Confusion matrix for random forest dataset 1

Adaboost Algorithm

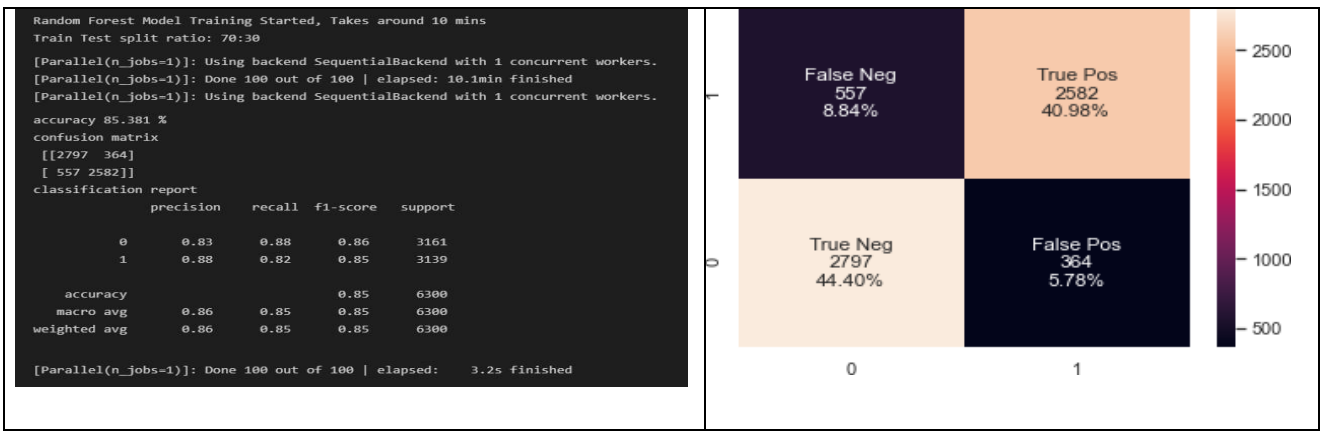
Dataset 1: Amazon labelled review dataset



Results:

Random forest Algorithm

Dataset 1: Amazon Labelled Reviews Dataset:



Random forest accuracy dataset 1

Confusion matrix for random forest dataset 1

Adaboost Algorithm

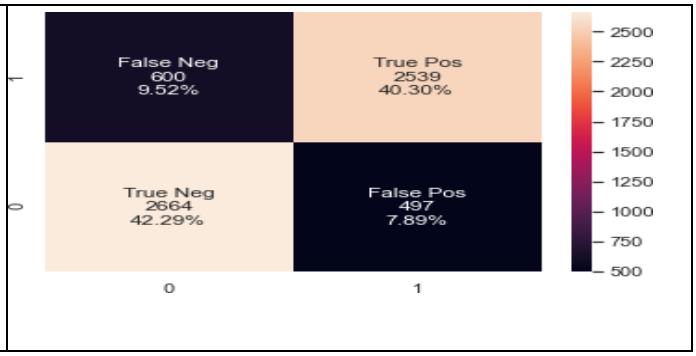
Dataset 1: Amazon Labelled Reviews Dataset:

Adaboost accuracy dataset 1

```
AdaBoost Model Training Started, Takes around 25 mins
accuracy 82.5873 %
confusion matrix
 [[2664  497]
 [ 600 2539]]
classification report
              precision    recall  f1-score   support

           0       0.82      0.84      0.83      3161
           1       0.84      0.81      0.82      3139

    accuracy                           0.83      6300
   macro avg       0.83      0.83      0.83      6300
weighted avg       0.83      0.83      0.83      6300
```



Confusion matrix for AdaBoost dataset 1

Dataset 2: After labeling dataset 2 we have the number of fake and real reviews as shown in figure.

Dataset 2 after labeling

```
No. of Fake Reviews:  1815
No. of Genuine Reviews:  1238
Fake to Genuine Review Ratio: 147.0 %
Total Fake Review Percentage: 59.0 %
```

Random forest Algorithm

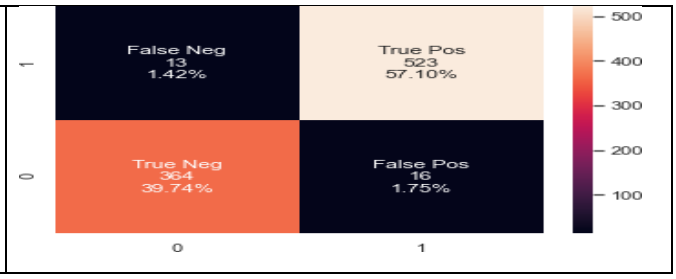
Dataset 2: Amazon Labelled Image Dataset**:**

```
Random Forest Model Training Started, Takes around 10 mins
Train Test split ratio: 70:30
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
accuracy 96.8341 %
confusion matrix
 [[364  16]
 [ 13 523]]
classification report
              precision    recall  f1-score   support

           0       0.97      0.96      0.96       380
           1       0.97      0.98      0.97       536

    accuracy                           0.97       916
   macro avg       0.97      0.97      0.97       916
weighted avg       0.97      0.97      0.97       916

[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    6.7s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    0.0s finished
```



Random forest accuracy dataset 2

Confusion matrix for random forest dataset 2

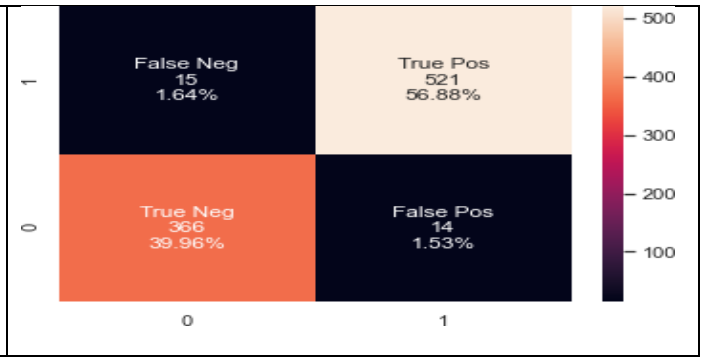
Adaboost Algorithm

Dataset 2: Amazon Labelled Image Dataset:

```
AdaBoost Model Training Started, Takes around 1 mins
Train Test split ratio: 70:30
accuracy 96.8341 %
confusion matrix
 [[366  14]
 [ 15 521]]
classification report
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       380
           1       0.97      0.97      0.97       536

    accuracy                           0.97       916
   macro avg       0.97      0.97      0.97       916
weighted avg       0.97      0.97      0.97       916
```



Adaboost accuracy dataset 2
Confusion matrix for AdaBoost dataset 2
Naïve Bayes Algorithm
Dataset 2: Amazon Labeled Image Dataset**:**
Naïve bayes accuracy dataset 2

```
Training Naive Bayes. It take huge amount of Memory and very Long Time
Training Naivye Bayes
Train Test split ratio: 70:30
accuracy 75.5459 %
confusion matrix
 [[257 123]
 [101 435]]
classification report
              precision    recall  f1-score   support

           0       0.72      0.68      0.70       380
           1       0.78      0.81      0.80       536

    accuracy                           0.76       916
   macro avg       0.75      0.74      0.75       916
weighted avg       0.75      0.76      0.75       916
```

Result Validation:

```
Enter Product Average Rating (1-5): 3
Enter Product Rating given by Reviewer (1-5): 2
Enter Review Title in English: Expensive
Enter Review Text in English: Not satisfied
Enter Product Category (ex: Electronics, Wireless, Watches etc.): wireless
Enter Product ID (Leave Empty if Unknown):
Enter Product Name: Earphones
Is the Purchase verified? (YES / NO): yes
Product Image provided by Reviewer? (YES / NO): yes
(1, 38014)
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    0.0s finished

Review Detected as Genuine by ML Model.
The reviewer provided Image, The review is more likely to be Genuine.
##############################
```

Conclusion

In this paper, we have developed a model aimed at distinguishing between authentic and fraudulent reviews by utilizing contextual and behavioral features. In addition to the content of the reviews, we have introduced a range of review-centric features to aid in the classification of fake reviews. Initially,

we trained the classifier on a labeled dataset (Dataset 1), which subsequently enabled us to generate labels for an unlabeled dataset (Dataset 2) that includes additional features. A notable review-centric feature proposed in this study is the "review image." Our findings indicate that incorporating the "review image" as a feature significantly enhances the classification of fake reviews; for instance, if a review is from a non-verified purchase but includes an image of the product, it may be deemed genuine when considered alongside other features. The work presented in this paper serves as a foundation for further exploration in the realm of fake review detection through various combinations of features. This thesis may prove beneficial for future researchers seeking to enhance fake review detection systems by leveraging the review image feature, as well as employing classifiers such as random forest and AdaBoost. The introduction of the "review image" feature represents a significant contribution of our research to the field of fake review classification. Furthermore, non-verified purchase reviews can also be accurately classified using this feature, thereby protecting both consumers and business owners from online fraud.

References

1. Nitin Jindal and Bing Liu. Review spam detection. In Proceedings of the 16th international conference on World Wide Web, pages 1189–1190. ACM, 2007.
2. Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. Distributional footprints of deceptive product reviews. ICWSM, 12:98–105, 2012.
3. Mukherjee,A.,Liu,B.,Glance,N. Spotting fake reviewer groups in consumer reviews. WWW. 2012.
4. Newman, M.L.,Pennebaker,J.W.,Berry,D.S.,Richards,J.M. Lying words: predicting deception from linguistic styles, Personality and Social Psychology Bulletin 29, 665–675. 2003.
5. Jin,X.,Lin,C.X.,Luo,J.,Han,J. SocialSpamGuard: A Data MiningBased Spam Detection System for Social Media Networks. PVLDB 4(12): 1458-1461 (2011).
6. Moghaddam,S.,Jamali,M.,Ester,M. ETF: extended tensor factorization model for personalizing prediction of review helpfulness. WSDM 2012: 163-172.
7. Saito,H.,Toyoda,M.,Kitsuregawa,M.,Aihara K. A Large-Scale Study of Link Spam Detection by Graph Algorithms (S). AIRWeb 2007.
8. Song,Y.,Kolcz,A.,Giles,L.C. Better Naive Bayes classification for high-precision spam detection. Softw., Pract. Exper. 39(11): 1003- 1024 (2009).
9. Xie,S.,Wang,G.,Lin,S.,Yu,P.S. Review spam detection via temporal pattern discovery. KDD 2012: 823-831.